Backchannel behavior in child-caregiver conversations

by

Jing Liu

S1049411

MA degree programme in

Linguistics and Communication Sciences (research)

Paris, 2022

Supervisor: Dr. Abdellah Fourtassi

Co-supervisor/Assessor: Dr. Asli Özyürek

Radboud University

Abstract

Conversation is a coordinative activity (Clark, 1996) that requires cooperative social interaction between interlocutors. The coordination nature of conversations has been the hallmark of children's socio-cognitive development as it involves the sophisticated ability to manage the flow of conversation through backchanneling i.e., signaling listener's attention through verbal (short responses like *Yeah*) and non-verbal behaviors (e.g. smiling, nodding). Previous studies on middle childhood children's backchannel behaviors were scarce and either conducted in highly controlled experimental settings (Hess and Johnson, 1988) or qualitative observation (Bodur et al., 2022). This thesis investigated middle-childhood children's production and responses of multimodal cues to elicit backchannels in naturalistic child-caregiver conversations (ChiCo: Bodur et al., 2021). In order to quantitatively infer the potential combinations of multimodal cues, two backchannel opportunity prediction models (Support Vector Machine model: *SVM*; Long short-term memory model: *LSTM*) were first trained respectively on children and adults' responses to the speaker cues in child-caregiver and adult-adult conversations and then tested on different input feature modalities. The comparable model performance between children and adults indicated that children between the ages of 6 and 12 can produce and respond to backchannel inviting cues as consistently as adults. The comparison of model input modalities suggested that features from vocal modality contributed most to backchannel occurrences, which may be caused by characteristics of child-directed speech and caregivers' linguistic alignment to scaffold children's language processing. In a word, although school-age children are still at the stage of developing socio-cognitive competencies, their performance of producing and responding to BC inviting cues are strikingly close to adult-level mastery. The broader impact of this thesis lies in the application of machine learning models in balancing the needs of ecologically valid settings and quantitative analysis.

Keywords: cognitive development; language acquisition; conversation; backchannel; nonverbal

Acknowledgements

The two-year ResMA journey was short but full of adventures. During this journey, one of the greatest experiences is to do an internship supervised by Prof. Abdellah Fourtassi and Prof. Asli Özyürek. I would like to express my deepest appreciation to Abdellah, who has supported and guided me throughout the internship period with his patience and philosophy of how to conduct research. Half a year ago, when I was groping in the maze, depressed and anxious about future career path, it was this internship opportunity that made me feel supported, slow down to sip the joy of exploration and cherish the moment of making progress in research. I would like to thank my co-supervisor Prof. Asli Özyürek, for the wonderful opportunity to attend and present my progress in the lab meetings. Without her invaluable guidance and feedback throughout my internship, this thesis would not have been possible. They are my role models who have rekindled my passion for research and made me determine to pursue a career in academia, no matter how hard it would be.

A very special gratitude goes out to all my colleagues to make this journey memorable. It has been a dear experience being surrounded by all these passionate researchers during the internship. I am grateful to Mitja for all the encouragement and unconditional help in formulating research questions, designing model structures and interpreting results. And my special thanks to Kübra, whose patience, support and wonderful work on Chico corpus have set a solid basis for the subsequent thesis work. With a special mention to Williams for his inspiration on forced alignment; Biswesh, Alafate and Mayank for their wonderful feedback.

I would like to thank the Honors Academy Programme for the support and inspiration of conducting research abroad. My sincere gratitude goes to Heleen, who inspired me to shape my research ideas to have more impact on society.

Finally, my heartfelt thanks to my parents whose sacrifices, love and encouragement helped me to tackle emotional ups and downs, uncertainty and doubt when I gave up the economically stable job three years ago to pursue my dream. Never had I expected my father, a traditional electrical engineer, would start to learn artificial intelligence only to discuss model structure with me, just like three years ago when he started to get familiar with psycholinguistic experiments. I am grateful to them for making me who I am today, for all. At this point, I owe you my deepest gratitude and love for your dedication and support.

List of figures

List of tables

Table of Contents

Chapter 1. Introduction

Conversation is a collective activity in which the joint goal and the mutual understanding are established between interlocutors through grounding (Clark and Schaefer, 1989). This process not only requires the speaker to keep track of the common ground but also the listener to send signals to indicate that they have received and perceived the other's signals (Levinson, 1979). Instead of taking the whole turn to explicitly manage the flow of conversation (e.g. the listener takes the turn **only** to express "I understand what you have said. Please continue."), interlocutors tend to apply some mechanisms to subtly enable mutual understanding (Clark and Wilkes-Gibbs, 1986) through backchannels, conversational repair and interactive alignment (Bangerter and Clark, 2003; Clark and Brennan, 1991; Dale et al., 2013; Fusaroli et al., 2017; Mills, Groningen, and Redeker, 2017; Pickering and Garrod, 2004).

As one of the mechanisms to achieve common ground, backchannels (hereafter BC) are assumed as listeners' subtle responses to signal understanding or agreement such as "*yes*" and "uh-huh" and/or non-verbal cues such as head nods and eyebrow movements (Bangerter and Clark, 2003; Schegloff, 1982; Yngve, 1970). Despite not having a narrative content, BCs are crucial for providing information concerning the quality of communication as they include information about perceptual processing, interpretation, evaluation and dispatch (fulfillment of a request, carrying out a command) (Bunt, 1994). In the collaborative process, speakers do not passively wait for BC responses, but emit cues (BC inviting cues) to confirm listeners' understanding through changes in prosody, gaze patterns, and other behaviors (Ward and Tsukahara, 2000; Lee et al., 2017).

While there is a large body of research investigating children's acquisition of linguistic structures (e.g., Kuhl, 2004), in comparison little is known about the development of language use in negotiating shared understanding with the interlocutor, especially in terms of inviting cues to elicit BC responses. The scarcity of studies on BC inviting cues in child-caregiver conversation can be attributed largely to methodological limitations. Backchannel is inherently spontaneous and relies on collaborative multimodal signaling such as eye gaze, head movement, intonation change and so on. This complex process makes it difficult to study using traditional research methods in language acquisition, whether experimental or observational. On the one hand, the experimental study (Hess and Johnson, 1988) on multimodal cues to elicit middle-childhood children's BC responses were conducted in highly controlled lab settings due to the constraint of data collection and analysis. While such in-lab experiments were quantitative and precise, the experimental paradigm of mechanically controlled speech production and instruction-oriented interactions might not reflect children's backchanneling skills in natural and spontaneous conversations. What's more, only a few variables were investigated in these studies, and more importantly, did not account for how isolated components manifested and interacted in naturalistic social interactions. On the other hand, the observational study (Bodur et al. 2022) have studied children's BC responses in the more ecologically valid interaction. Notably, it generally involved qualitative analyses, and the degree to which children can combine multiple cues to contribute to the effective collaboration in a conversation is not well known.

Facilitated by the need of quantitative inferences of BC inviting cues over more ecologically valid multimodal conversational data, this thesis leverages recent technology development from *Natural Language Processing* to examine children's BC skills using more complex machine learning models due to the models' capacity to process more ecologically valid multimodal data and the potential to interpret factors by manipulating model input.

Besides advancing our fundamental understanding of conversational development, this work also facilitates development of child-centered conversational agents. Prior studies have shown that children tend to interact with robot partners in a human-like manner. For example, they are sensitive to verbal and non-verbal signals, such as eye gaze (Okumura et al., 2013), and often attribute mentalistic competencies to the robot (Marchetti et al., 2018). In this respect, BC generation towards children's multimodal cues is associated to a greater interactional potential of human-like behavior to design more naturalistic conversational agents.

This thesis aims to investigate the contribution of different (combinations of) speaker cues that middle-childhood children employ in natural interactions using machine learning models. The research questions will be elaborated in Section 2.5

The rest of this thesis is organized as follows:

• Chapter 2 commences with a review on related work on BC responses from perspectives of communicative functions, potential multimodal eliciting cues, middle-childhood children's development and computational models in order to provide a theoretical and methodological backdrop to the thesis.

• Chapter 3 introduces the dataset and pre-processing steps to set a basis for the following studies.

• In Chapter 4, a support vector machine model (SVM) is introduced to investigate the prospective inviting cues.

• Chapter 5 applies long short-term memory neural networks (LSTM) to incorporate sequential information in prediction.

• Chapter 6 provides a conclusion to the thesis, drawing all the discussions to a close. The main aims and objectives are revisited and ways in which these have been met are clearly outlined. Furthermore, the limitations of the study are explored and how these limitations might be overcome in future studies is considered.

Chapter 2. Related work

This chapter lays the foundations for the exploration of BC inviting cues and children's cognitive ability. Section 2.1 elaborated the inherent relationship between BC and common ground based on the wealth of previous research. Section 2.2 highlighted potential multimodal cues to elicit children's BC responses. Section 2.3 reviewed previous research on BC behaviors of middle-childhood children and raised hypotheses regarding the research questions. Section 2.4 reviewed previous studies on computational models in predicting BC behaviors. Section 2.5 raised research questions in this thesis.

2.1 BC and common ground

Effective communication calls for the success in building mutual understanding between interlocutors, the process of which was defined as grounding: a constant evaluation of whether I share mutual beliefs, knowledge, and understanding sufficient for the purpose of the situation (Clark and Brennan, 1991). A diverse set of disciplines have approached this issue, from psycholinguistics to conversation analysis, highlighting several conversational strategies for coordinating interactions. Interlocutors might ensure common ground by subtly confirming their understanding (backchanneling), more explicitly signaling misunderstanding and correcting each other (conversational repair), or by re-using each other's linguistic forms (linguistic alignment).

In contrast to the "main channel", through which the speaker emits his/her information, over the backchannel (BC hereafter), the listener provides feedback without claiming the floor (White,1989). Therefore, the term 'backchannel' is also referred as 'accompaniment signals' (Kendon, 1967), 'receipt tokens' (Heritage, 1984), 'hearer signals' (Bublitz, 1988), 'minimal responses' (Fellegy, 1995) and 'reactive tokens' (Clancy et al., 1996) in previous studies. Nevertheless, BC is considered as the listener's feedback to signal attention without interrupting the flow of conversation.

BCs are multimodal in nature (White, 1989). Verbal BCs were divided into three types: 'simple' (raised by Oreström, 1983), 'double' and 'complex' (Oreström, 1983: 121;Tottie, 1991: 263) according to the constituting lexicons. Simple forms consisted of brief "mono or disyllabic utterances" (Gardner, 2001: 14) like *yeah mmm*. Double BCs, indicated by the name, comprise a sequence of repeated lexicons such as *yeah yeah*. Complex backchannels moved from lexicon level into phrasal or even clause level, which were defined as multiple-word chunks composed of more than one "open-class lexical items" (Tottie, 1991), such as *yeah I know*. Notably, complex backchannels were assumed to function beyond showing understanding, but rather signaling a desire to take the floor in the near future similar to "a raised hand in a classroom" (Dittman and Llewellyn,1968; Oreström, 1983: 124). Some researchers even hypothesized that complex vocal backchannels can be converted into a turn on the condition that "current speaker shows no willingness to continue speaking" (Cutrone, 2005: 242; Pipek, 2007). Apart from verbal BCs, non-verbal BCs are also non-negligible given prosody, gestures, facial expression and body movement are usually intertwined (Goffman, 1967) and integrated by interlocutors. For instance, it is quite frequent to see someone Nod his head or using phrases like "okay" and "uh-huh" when another person is talking.

BC was considered as one of the mechanisms to achieve common ground. Previous literature categorized the functions as four types (O'Keeffe and Adolphs, 2008: 84): i.) *Continuers* as floor-yielding signals to show the addressee' attention and desire to maintain the speaker's floor; ii.*) Convergence markers* to reinforce mutual understanding throughout the discourse by showing agreement; iii.) *Engaged response* to emit emotive signals such as surprise, shock, sympathy and so on; iv.) *Information receipt signal* to indicate the close or shift of a topic. Another trend of research summarized BC functions in the following two aspects: reinforcing Grice's Maxim of Cooperation in communication (1989) as 'non-floor holding devices' (O'Keeffe and Adolphs, 2008: 74); or marking convergence (Watzlawick et al., 1967); that is, functioning both organizationally and relationally in discourse (O'Keeffe and Adolphs, 2008: 87). Accordingly, BC transmitted by the listener is categorized as "generic" and "specific" based on the two aspects above (Bavelas et al., 2000). While the generic BC are used for indicating comprehension and attention to sustain the conversation flow without responding to the narrative content of the moment (Schegloff, 1982; Goodwin, 1986; Stivers, 2008), specific BC is closely related to what the speaker says and does (Goodwin, 1986; Bavelas and Gerwing, 2011). As highlighted by Knudsen et al.(2020), generic BC encourages the production of new information while specific BC provides evaluations or attitudes of previous information in the conversation.

2.2 Multimodal cues to elicit BC responses

Information transmission in vocal and visual channels such as prosody, gestures and facial expressions are usually intertwined with verbal content, especially in face-to-face conversation scenarios (Morgenstern, 2014). Joint cues (e.g. simultaneously making eye-contact while raising intonation) have been found to quadratically increase the likelihood of eliciting a backchannel response from listeners (Hess,1988). However, it is still unclear whether children and adults pay equally weight for these cues given their protracted development of communicative ability. Therefore, this section raised potential multimodal cues to elicit children's BC responses based on previous adult BC prediction studies to set a basis for the input features in our modeling study (see Table 1 in the appendix for a summary).

2.2.1 Verbal modality

a. Part-Of-Speech (POS)

POS is considered as the indicator of the utterance ends, especially for some languages with fixed structures. The discourse markers are assumed to be the sign of discourse organization, often associated with transition between discourse units, which may elicit communicative responses. This has been validated by many BC prediction studies (Cathcart et al.,2003; Truong et al., 2010; Boudin et al., 2021). For instance, Cathcart et al. (2003) developed a shallow BC prediction model based on pause duration and the POS tags of the preceding words.

b. Lexical information

Lexico-semantic information was used as another potential factor to elicit BC feedback (Boudin et al., 2021). Specifically, the word polarity (positive, negative) and aspect (concreteness) obtained from word lists (Bonin et al., 2018) were integrated in a logistic regression model given that information can be associated to specific listener's reaction on a certain level of emotion, but also the discourse referent associated to concrete words. Notably, this feature is limited due to its decontextualized nature.

c. Information entropy

Compared with the decontextualized lexical information, information entropy integrates the context information in the form of conditional probability. It can be another potential factor given that the grounding process can be reflected by the converging trend of information density of different speaker roles, i.e. “topic initiator” and “topic responder” (Xu and Reitter, 2018). Thus, it is hypothesized that BC, as the mechanism for common ground, occurs as the converging trend develops. More specifically, at the beginning of the conversation, when the responder at first knows little about the new topic; the purpose of his/her early utterances, is to let the initiator know that s/he has received the new information, which is hypothesized as the typical places where simpler utterances containing less lexical information occur, such as short acknowledging BC utterances, and short comments. However, there is no empirical evidence on its predictive effect yet.

d. Word embedding

Closely related with lexical information obtained from the word list, word embeddings reveal more concrete word information by decomposing and projecting the semantic information in a highly dimensional space. can be another predictor as an encoding of the speakers’ word history. Ruede et al (2017) tested the efficiency of word embedding by adding Word2Vec to prosodic-based LSTM model, which increased the f1 score to 0.39. Ortega et al. (2020) also used Word2Vec and reported the performance level of about 58% accuracy on three types of BC categories, i.e. continuer, assessment, and nonBC.

e. Verbose

Wordy, a long contiguous utterance, was found to have a predictive effect on children’s BC responses (Park et al., 2017; Gravano and J. Hirschberg). Instead of directly computing the number words within an utterance, Park et al. (2017) predicted BC opportunity based on inter-pausal units (IPUs).

2.2.2 Vocal modality

a. Pitch

The fundamental frequency has been proved as one of the most salient features for BC opportunities (Sugito, 1994; Ward and Tsukahara, 2000) based on previous observations that BC frequently occur at junctures between phonemic clauses (Dittmann and Llewellyn, 1967), and at "the ends of intonation units with non-final intonation contours" (Clancy et al., 1996), where the declining pitch slope serves to foreshadow these. To be more specific, four related features are involved in terms of BC prediction: low pitch region, pitch slope, pitch variation and pitch absolute value. This is also generally related with the following communicative functions.

First, declination or boundary tones often occur at points where the speaker considers that s/he has transmitted enough information for the listener to infer the speaker's point (Ward and Tsukahara, 2000), which are not necessarily at the end of utterance. Therefore, BC sometimes even appears before the speaker has completed a grammatical phrase or full proposition. Occasionally, a low pitch region marks the repetition of previous words, though produced for emphasis or clarity. Such cases can be interpreted as conveying 'I said it again, did you get it that time?' (Ward and Tsukahara, 2000: Page 1186).

Second, some pathological studies have indicated that low pitch region is one of the most characterizing features for disfluencies and formulation difficulties, especially in English (Ward, 1999; Xue et al., 2021) and Dutch (Van Bemmel et al., 2021). A related communicative function is taking the floor before actually saying anything, where the speaker utters some fillers to call for attention, or, which are typically in low pitch regions. In this case, such cues can be interpreted as “I'm stuck, but keep listening, something meaningful will come out soon” (Ward and Tsukahara, 2000: Page 1186).

Third, some special utterances like sentence-final particles typically functioning as “agreement seeking” or 'invite collaboration' (Cook, 1992), like 'you know', which is known to appear with low pitch regions and associated with several turn-taking operations (Tanaka, in press).

b. Lengthened vowels

Cooccurring with the low pitch region, vowel lengthening has also been found to precede BC responses, especially in cases of disfluencies and agreement-seeking tokens (Maynard, 1989; Ward and Tsukahara, 2000). Indeed, lengthening is assumed as a consequence of producing a low pitch region of sufficient length, in those cases where there is only a single syllable of lexical content to work with, for example '嗯' (pronunciation: en4; meaning: yes) in Mandarin Chinese. This hypothesis is supported by the fact that lengthening seems to occur less often when the low pitch region falls on longer words and phrases.

c. Voice quality

Co-occurring with some pitch features (Kuang,2017), creaky voice feature, originally applied in pathological studies (e.g. Almeida, 2010; Merkus et al, 2020), have been transferred to BC prediction (Gravano and Hirschberg, 2009). Acoustically, voice quality is characterized by jitter (average absolute variations between pitch consecutive period), shimmer (average absolute variations in pulse amplitude) and irregular F0 measured by Harmonic-to-Noise Ratios (HNR) (de Krom, 1993) as the result of a tightening of the vocal folds that prohibit longer and continuous vibrate (Ladefoged 2006, Ladefoged and Maddieson 1996). In the context of conversation scenarios, creaky voice is generally considered to be related with hesitation and disfluency, thus related with BC occurrence. This has been validated in previous BC studies (Gravano and Hirschberg, 2009; Levitan et al., 2011).

d. Loudness

Some energy-related features like intensity and loudness have also been found to be predictors for BC opportunities (Ruede et al., 2017). Like pitch, some pathological and L2 speech studies have indicated that loudness marked disfluencies and formulation difficulties (Liu and Strike, 2022; Presenti et al., 2022; Van Bemmel et al., 2021). The presence of fillers, which are generally in a lower volume due to speech reduction, can be the possible indicator of the BC occurrences. What’s more, some L2 speech production studies indicate that confidence is reflected by loudness during articulation (Cuchiarinni et al., 2000; Liu and Strik, 2022). Therefore, the declining loudness may reflect the need of listeners’ BC responses as the support of continuing speech production.

e. MFCCs

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively constitute Mel-frequency cepstrum (MFC), which represents the short-term power spectrum of a sound. Conventional ASR studies tend to extract 12 cepstra plus log-energy and their first-order time derivatives as the input for training considering that the lower order coefficients already contain most of the information about the overall spectral shape. MFCCs have been widely used in BC prediction studies (Goswami et al., 2020; Jain et al., 2021; Murry et al., 2021; Ruede et al., 2017). In particular, Murry et al. (2021) augmented the MFCC features by warping them across time and masking blocks of consecutive frequency channels, which improved the model’s robustness to partial loss of information and deformations across time.

f. Pause

Hesitations occur frequently in everyday conversations for a wide range of reasons including: lexical access, structuring of utterances, and requesting feedback from the listener (Carlson et al., 2006). Pause has been used in a lot of BC prediction studies (Boudin et al., 2021; Cathcar et al., 2003; Jain et al., 2021; Park et al., 2017).

2.2.3 Visual modality

a. Eye gaze

Eye gaze has been assumed to influence the flow of conversation by signaling to the listener that the speaker is waiting for feedback (Morency et al., 2009) or showing that the addressee is attending to the speaker (Clark, 1989,1996; Nakano et al., 2003; Lee et al., 2017). Prosodic cues such as pauses in speech or changes in pitch have been found to be too subtle for young children to acknowledge, but their cueing context can be strengthened by adding multimodal behaviors like a gaze cue (Park et al., 2017). This is validated in previous BC prediction studies, which either used manual annotation on speakers' binary gaze direction (looking at the listener or not) (Bodur et al., 2022; Morency et al., 2009; Nakano et al., 2003) or automatic extraction of related features (*OpenFace:* Baltrušaitis et al.,2016) like velocity and acceleration of eye movements as well as blink rate and pupil dilation (Goswami et al., 2020; Jain et al., 2021).

b. Head movements

Similar to eye gaze, head movements, as one of the contributors to lower levels of grounding (Nakano et al., 2003), tend to signify interlocutors' access to each other's communicative actions, like showing continuation of contact, perception and understanding (BC continuers), showing agreement to the speaker's content (convergence marker) or eliciting feedbacks from the listeners (Cerrato, 2007). In particular, speakers' Nod has shown concurrent patterns with phonetic patterning, lexical forms, semantic content (Blache et al., 2008; Cerrato and Skhiri, 2003; Kendon, 1972; McClave, 2000; Goldin-Meadow, 1999; Kendon, 1972), thus synergically signaling speakers' intent of inquiring understanding.

Recent advancement of technology has made the automatic annotation of head movements possible. Previous studies either use features such as rotational velocity and acceleration rate captured by camera (Jain et al, 2021; Jindal et al, 2020); or spatial information captured from a laser tracker to annotate head movements (Michalowski et al., 2006).

c. Facial expressions

Facial expressions such as eyebrow movements, smiles and laughs are a common source of back-channel communication as they can exhibit listener's surprise, confusion and understanding, which are generally synergized by other gestures (Brunner, 1979). For instance, Dittmann and Llewellyn (1968) pointed out that the simultaneous production of a short smile with a head Nod was a typical signal of attention. However, no studies have been conducted on the BC inviting cues yet.

d. Posture

Posture movement like leaning forward or backward is another source of signaling attention and inviting feedback. As indicated by Park (2017), leaning towards the speaker is considered as a positive response to the speech content whereas leaning backward may occur as a turn exchange or BC inviting cues. However, few BC prediction studies have integrated this cue due to the detection and implementation difficulty.

2.2.4 Modality comparison and integration

Notably, most studies that integrated the above-mentioned multimodal cues to predict adults' BCs are engineer-oriented, thus ignoring the interpretability of the input cues and potential mechanisms. Only a few studies have examined the contribution of different features or modalities.

For instance, Ruede et al. (2017) tested different combinations of features in LSTM models using ablation experiments. Their results indicated that semantic features, i.e., Word2Vec, together with some prosodic features (FFV, MFCCs and absolute pitch value) performed best. Goswami et al. (2020) adapted Partial Dependency Plots (PDPs) algorithm to investigate how some important features influence BC occurrence. Their results showed that f_0, pupil dilation and MFCCs ranked highest among all the visual and vocal features. Boudin et al. (2021) fitted a logistic regression model to compare the contribution of different modalities based on the coefficients and accuracy of different modalities. Their results indicate that vocal modality has reached the highest accuracy, followed by visual and verbal modalities. Nevertheless, these studies indicate that adults benefit from cross-modal BC cues.

2.3 BC development

Spoken language production has been claimed to change across the lifespan (Mortensen, et al., 2006). To exemplify, child-directed speech is clearer, higher in pitch, and slower in speed (Peccei,1999) and has more pauses and distinct pronunciation with exaggerated intonation. Children also use pet names, simple sentences, repetition, tag questions, and baby talk words more frequently. Accordingly, such production differences may exert an influence in BC inviting cues.

BC behavior was found to differ among age groups (Geertzen, 2015; Wong and Kruger, 2018). In a comparative study of 2- to 5-year-old children's BC production, elder children were found to spend more time Nod heads, smiling and gazing at adult speakers, suggesting that they better understand a listener's role in providing collaborative feedback (Miller et al., 1985). Another comparison of the adult age group reflected younger adults' higher frequency of BC production than older speakers in both task-oriented dialogues (MapTask: Kemper et al., 1998) or simultaneous conversations (Gould and Dixon,1993). This difference has been explained as an increased "willingness and ability to take on the cognitively demanding task of dividing one's attention between monitoring the social situation and planning one's own speech productions" (Gould and Dixon, 1993).

Although there is extensive research on early-childhood and adult BC behaviors, limited prior work exists in investigating middle-childhood children's BC responses, especially their response to inviting cues. For instance, BC frequency was found to increase with age for children between 7 to 12 years old (Dittmann, 1972; Hess, 1988). A further analysis suggested that joint cues quadratically increased the likelihood of eliciting a backchannel response (Hess and Johnston, 1988; Gravano and Hirschberg, 2009; Lee et al., 2017). However, these studies were questioned either in terms of the relatively small sample size or the highly controlled experimental settings. In Dittmann's study (1972), only 6 child-adult conversations in a laboratory setting were analyzed, resulting in only one sample per age group. Realizing such deficiency, the following study (Hess, 1988) increased the sample size but utilized the predesignated instructional interaction with fixed length of clause boundaries, pauses that lasted more than four-tenths of a second and speaker eye gaze toward the listener. Notably, the highly controlled environment and nature of instructional language typically requires listeners' longer tolerance for ambiguity or partial understanding compared with daily conversation, which might not reflect children's BC skills in natural and spontaneous conversations. A recent study by Bodur et al. (2022) turned to more natural and spontaneous settings using semi-structured conversations and compared (different types of) BC responses in child-caregiver and adult-adult conversations. In contrast with previous findings (Dittmann, 1972; Hess & Johnston, 1988), their results revealed that children produced BC at a similar rate than adults in family dyads. A further examination of speaker cues (defined as speech, gaze and short pause) demonstrated that BC distribution was largely similar between children-caregiver and adult-adult conversations in a family context. Notably, such analysis was qualitative and the extent to which children can combine multiple cues to contribute to the effective collaboration in a conversation is not well known for the moment. In this work, I propose to instantiate quantitative analysis using machine learning models that can learn from naturalistic data.

2.4 BC prediction models

In the past few years, the research community in dialogue system has shown a keen interest in modeling the listener's backchanneling behavior.

Earlier models of identifying BC opportunities generally exploited hand-crafted rules abstracted on the linguistic level due to the constraint of model structure (e.g. Moubayed et al., 2009; Park et al.,2017; Ruede et al., 2017; Tuong et al., 2011; Ward and Tsukahara, 2000). Denny (1985) was the first to describe BC cues based on speaker's intonation, mutual gaze, gesture and "filled pauses" such as *mm-hmm*, and grammatical completion in the preceding context. Notably, most of the features largely relied on manual annotations, which truncated a considerable number of details; for instance, pitch variations were simply represented as increasing/decreasing slopes. Following their descriptive model, Koiso et al. (1998) expanded the feature set by adding verbal features like the preceding word's POS and more fine-grained acoustic predictors such as duration of the final phoneme, energy pattern, and energy peak. An intercorrelation analysis was first introduced in their study to explore the interaction of different modalities, which indicated that the predictive effect of POS was augmented by the co-occurring prosodic features. However, the duration of phoneme was skeptical considering the continuity of speech production, making the speech reduction phenomena quite common. Inspired by the intersection between verbal and acoustic modalities, Ward and Tsukahara (2000)'s model predicted BC occurrence whenever the speaker produced a region of low pitch lasting 110ms. This was based on the observation that such regions were often accompanied with grammatical completion, especially in English (Ward, 1999) and Japanese (Ward, 1996; 1997). However, their model only reached 34% accuracy in Japanese and 18% in English. Some recent studies (Boudin et al.,2021) integrated lexical-semantic information by manually annotating word polarity (positive, negative) and aspect (concreteness) based on the given word lists (Bonin et al., 2018). Considering the highly dependence of semantic meaning on context (Firth,1957), the simple binary features and context-deprived annotations may not reflect the contextual meaning. In a nutshell, these highly abstract features applied in these models, on the one hand, are conceptually clear and interpretable; but on the other hand, tend to yield low performance due to the unnecessary information reduction.

Recent advancement of data-driven models has yielded higher performance compared with hand-crafted models. Solorio et al. (2006) used locally weighted linear regression (LWLR) to predict BC opportunities with prosodic features. The instance-based learning classifies a query point by examining how similar points are in the training data, which is computed as a Euclidean distance. The weight given to a data point is proportional to how similar it is to the query point. Their results achieve as good performance as that obtained using a laboriously developed and predefined rule. Following study (Nishimura et al. ,2007) proposed a decision-tree approach for producing BCs based on prosodic features. The system analyzed speech in 100ms intervals and generates BCs as well as other paralinguistic cues (e.g., turn taking) as a function of pitch and power contours. Compared with the previous rule-based prosodic model (Ward and Tsukahara, 2000), the decision-tree model achieved comparable naturalness to that of human-human dialog timing as obtained from subsequent human evaluation on the model output. Considering the high inter-person variability in BC behaviors (Cathcart et al., 2003), Morency et al. (2010) demonstrated that sequential models based on an associated probability such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRF) significantly improved predictions of previous hand-crafted models by Ward and Tsukahara (2000). More recently, many researchers utilized deep learning techniques for predicting BCs (Jain et al, 2021; Murray et al., 2021; Ruede et al., 2017; Turker et al; 2018). Among these models, the Long short-term memory (LSTM) models are gaining popularity due to its prominence in dealing with sequences of data (such as speech or video). The input features are generally extracted from three modalities on the frame level and concatenated before fed into the model. The model performance has been improved to a large extent due to the structure strength. Model settings are summarized in Table 3 in the appendix.

2.5 Research questions

This thesis extends these prior works by examining the contribution of different (combinations of) speaker cues that middle-childhood children employ in natural interactions. Specifically, two research questions are raised.

Firstly, I ask how middle-childhood children's BC behavior compares to adult-level mastery in terms of their responses and productions of multimodal cues. Since the collaborative nature of backchannel is instantiated in a predictive framework, I ask to what extent children's BC behavior can be predicted by the model (compared to adults), that is, if children capitalize consistently on inviting cues from caregivers, which would lead to high predictive models of children as listeners. Then I investigate the role of each modality (verbal, vocal, visual) in this prediction as well as the integration of multiple modalities in predicting BC behaviors. It is hypothesized that children may exhibit lower consistency in producing and responding to BC inviting cues due to their protracted development of communicative ability, which can be reflected on the comparatively lower model performance trained on child-caregiver conversations.

Secondly, I ask how children's BC behavior differs in terms of its type (specific vs. generic). Considering their different communicative functions, the specific BC may be more predictable by the preceding speaker features as it is closely related to the speaker context.

In the following chapters, pre-processing steps were first introduced to extract and temporally align the selected multimodal features. Then two computational models were respectively constructed to compare children and adults' prediction performance as well as the influence of BC type. A support vector machine model (SVM) was first selected as a baseline model due to its strength in processing highly dimensional and no presupposition of feature independence. Based on a further inspection of input feature structure and model performance, I proposed to apply the long short-term memory neural networks (LSTM) to incorporate sequential structure of the input features in Chapter 5.

Chapter 3. Data and Methods

This chapter introduces the dataset and pre-processing steps to set a basis for the following analysis.

3.1 Dataset

The Child-Caregiver Interaction Corpus (ChiCo: Bodur et al., 2021) was used for analysis, which consisted of 349-minute zoom-coordinated conversations in two conditions: the child-caregiver conversation as the condition of interest, and the adult-adult conversation as the control condition the "end-state" that children should reach. Children were between 6 and 11-years old (one 6-year-old, two 7-year-olds, two 8-year-olds, three 9-year-olds, one 10-year-old and one 11-year-old; mean age: 8.7; SD = 1.37). All 10 children were native French speakers, 5 of whom were reported to be bilinguals speaking English/Portuguese/Spanish/Czech with the other caregiver. In adult-adult conversation, the same caregivers talked to adults following the same procedure (marked as adult 1 in the following analysis).

Each conversation was composed of three stages: first, the caregiver/adult 1 explained the task, then they started to play a word-guessing game, i.e. guessing the word in each other's mind, for around 10 minutes, and finally they initiated a more spontaneous conversation on their experience and suggestions of the game. After a word was guessed, the interlocutors changed their roles. Each recording contained three to four rounds of game.

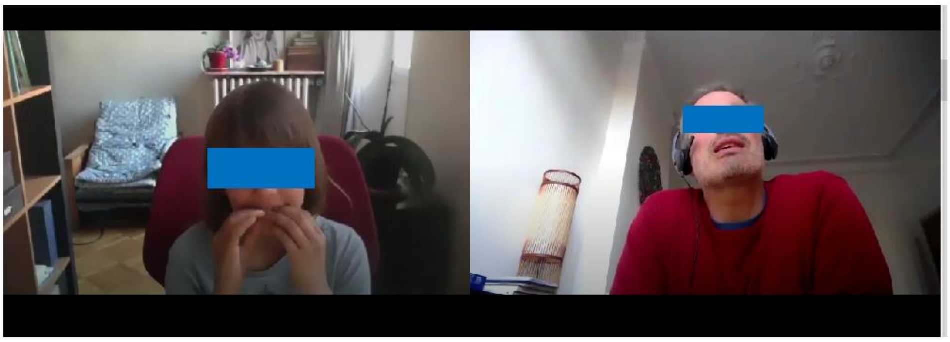


Figure 3.1 A snapshot of one of the recording sessions involving a child and her caregiver communicating through Zoom

The original dataset has been manually annotated in terms of (types of) BC responses and non-verbal behaviors (Bodur et al., 2022; see Table 2 and Table 3 in Appendix).

3.2 Feature extraction

Since the pipeline from a recording to the eventually selected features is quite extensive, it will be described in this section. As Figure 3.2 below shows, feature engineering pipeline consists of following steps:

1.) Noise reduction

2.) Forced alignment

3.) NonBC down-sampling

4.) Manual annotation of dialogue structure

5.) Feature extraction

6.) Feature selection

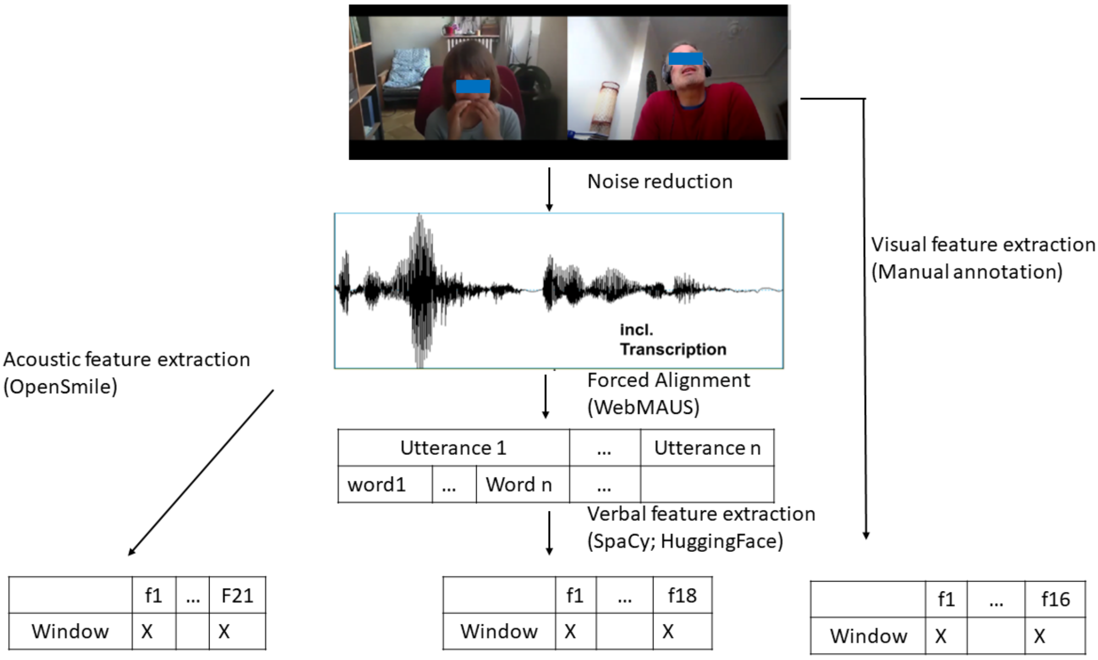


Figure 3.2 Pipeline of data pre-processing and feature extraction

3.2.1 Data preprocessing

Noise reduction

To reduce the influence network instability, I applied the non-stationary noise reduction algorithm to all the audios (Sainburg et al., 2020). This algorithm was motivated by Per-Channel Energy Normalization in bioacoustics which allowed the noise gate to change over time. The noise reduction typically went through the following steps:

1. Apply an IIR filter forward and backward on each frequency channel to smooth the spectrogram over time
2. Compute a mask based on the time-smoothed spectrogram
3. Apply a filter over frequency and time to smooth the mask
4. Apply the mask to the spectrogram

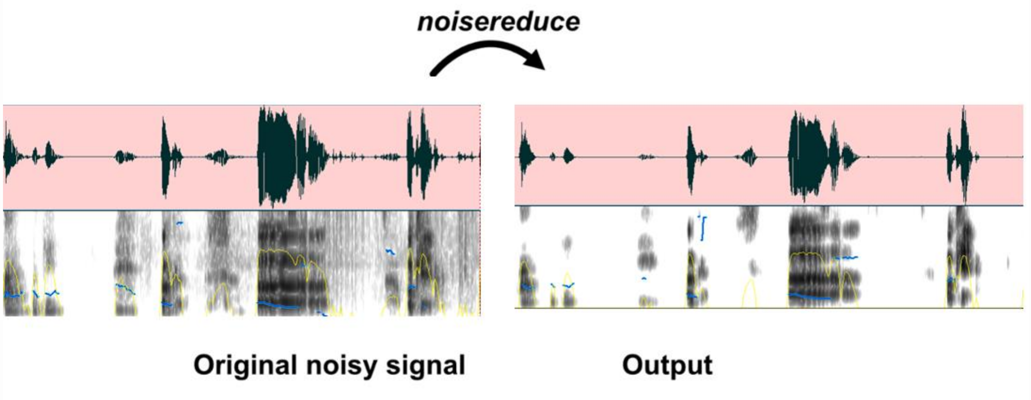


Figure 3.3 Visualization of noise reduction

Forced Alignment

To integrate features from different modalities, the forced alignment was conducted to automatically annotate time-aligned segmentation on the word level. Considering the overlapping speech and noise of conversation data, recording of each conversation was first split into two mono-channel audios containing only one speaker's voice and segmented at the utterance level based on the start points of each utterance in transcriptions. Then the utterance transcription and audios were sent to the Hidden Markov Models to obtain phonetic segments. The performance of three types of forced aligners on randomly sampled 50 utterances was compared (Montreal Forced Aligner: McAuliffe et al., 2017; SPPAS: Bigi and Christine, 2018; WebMAUS: Kisle et al., 2017). WebMAUS was then selected due to the highest accuracy.

3.2.2 Feature extraction

Before extracting multimodal features, the topic episodes and nonBC down-sampling were first conducted.

Manual annotation of dialogue structure

To compute the contextualized information entropy (Giulianelli and Fernández, 2021; Xu and Reitter, 2018), the topic episode was manually annotated based on the phase of the game and each round within the game of each conversation (see Table 3 in Appendix). The computation will be elaborated below.

NonBC down-sampling

As the original dataset only contained BC instances, the non-BC behaviors were sampled randomly and balanced on each speaker level as Table 3 shows. Following Jain et al. (2021)'s sampling procedure, non-BC behaviors were sampled when:

(1) the listener is not speaking, and

(2) the listener is not backchanneling in that time frame

Table 3.1

Overview of dataset

	BC occurrence			BC type		
	BC	**NonBC**	**Total**	**Generic**	**Specific**	**Total**
All	2841	2841	5682	1218	1623	2841
Children	573	573	1146	484	89	573
Caregiver	454	454	908	187	267	454
A1	810	810	1620	448	362	810
A2	1004	1004	2008	423	581	1004

Feature extraction

Following previous BC modeling studies (Jain et al., 2021; Jindal et al., 2020), speaker cues were extracted within the given context window (preceding 2 or 3 seconds). Two types of feature sets were extracted to be applied in models in Chapter 4 and Chapter 5 respectively: window-based and frame-based. While the former extracted features from the whole context window; the latter extracted features every 50ms to adapt to the sequential model structure (Roddy et al., 2018). These features were from visual, vocal and verbal modalities respectively as summarized in Table 4 below.

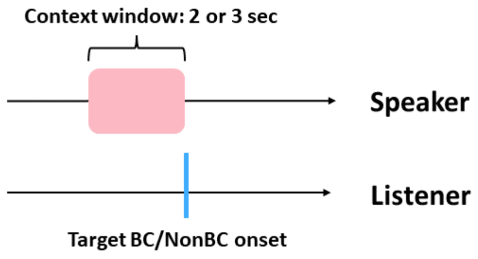


Figure 3.4 Visualization of the window-based feature extraction

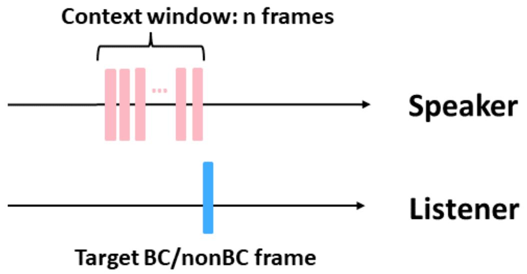


Figure 3.5 Visualization of the frame-based feature extraction

Visual features were extracted from manual annotations in the dataset (for details, see Table 2 in appendix). One-hot encodings were used in two feature sets, where each feature was "switched on" in the time window or a single frame of 50ms. Apart from one-hot encodings, duration-related features were added to the window-based feature set by extracting feature duration in the given context window.

I used the Opensmile (Eyben et al., 2010) toolkit to extract acoustic features. These features were subsets of eGeMAPS where potential cues were applied in previous studies (see Table 4 below). To minimize identify-confounding (Neto et al., 2019) and inter-speaker differences, data was standardized by calculating the z-scores on the speaker level, making each feature have a zero mean and a unit standard deviation:

$$\hat{x}_i = (x_i - \mu_i)/\emptyset_i$$

where xi, μ_i and $\emptyset$ are the original value, the mean, and the standard deviation of feature i respectively. This standardization method was inspired by *Cepstral Mean and Variance Normalization* (CMVN: Viikkiand Laurila, 1998) for noise robust speech recognition and Mandarin tone marking scheme (Five-level tone marking scheme: Chao, 1920) to remove gender interference in tone recognition (Liu, 2021). While the original methods were only applied on MFCCs and fundamental frequency, I expanded it into all the acoustic features.

POS tags were extracted through SpaCy (Honnibaland Montani,2017) toolkit, resulting in 17 types of tags altogether. For the window-based feature set, the number of occurrences of each POS type was added. For the frame-based feature set, one-hot encodings were used, where each POS tag was "switched on" for a given context window or single frame of 50ms. For those words longer than 50ms, the corresponding POS tag was added as long as part of the word occurred in the given time frame.

The word probabilities were extracted using DialoGPT (Zhang et al., 2021), an autoregressive Transformer language model trained and fine tuned on French dialogue materials extracted from films, interviews and theater plays, which provided more accurate probability estimates than the n-gram or GPT-2 models trained on passages (Genzel and Charniak, 2002, 2003; Doyle and Frank, 2015; Qian and Jaeger, 2011; Xu and Reitter, 2018; Giulianelli et al., 2021). I relied on HuggingFace's implementation of DialoGPT with default tokenizers and parameters (Wolf et al., 2020). In particular, three values related with word probabilities were computed.

Word surprisal was calculated on the last word appearing in the context window using the formula below:

$$\text{Surprisal} = -\log_2 P\ (\text{word} \mid \text{context})$$

The contextualized information content of an utterance is computed by averaging over the negative logarithms of all word probabilities, conditioned only on the preceding words in the same utterance:

$$\mathrm{H(U)} = -\frac{1}{|U|}\sum_{w_i \in U} \log_2\ P(w_i|w_1, \dots, w_{i-1})$$

The contextualized information content of a sentence was computed as the average per-word negative probability, conditioned on the preceding words in the sentence as well as on the entire relevant discourse context. Instead of exploiting a topic segmentation algorithm (Xu and Reitter, 2018), I manually segmented the contextual units by using the inherent (task-related) structure of task-oriented dialogues as elaborated in above.

$$\mathrm{H(U|C)} = -\frac{1}{|U|}\sum_{w_i \in U} \log_2\ P(w_i|w_1, \dots, w_{i-1}, C)$$

3.2.3 Feature selection

Most of the features were selected based on previous adult BC prediction studies. I additionally tested the predictive effects of word probability related features: word surprisal, decontextualized information entropy H(U) and contextualized information entropy as they have not been applied in previous studies yet. Specifically, a logistic regression model was fitted with the occurrence of BC as the response variables and information entropy, as well as interaction between conversation type (child-caregiver v.s. adult-adult) as the predictors, specified as *BC occurrence ~ entropy(or surprisal) * conversation type + (1 | interlocutor)*. As a result, the word surprisal was shown to have a significant predictive effect for BC occurrence (β = 0.447, p < 0.001) whereas I didn't find the significant predictive effect of H(U) or H(U|C) on BC occurrence (H(U): β = -0.020, p = 0.498; H(U|C): β = 0.017, p = 0.557). Therefore, contextualized and decontextualized entropy were removed from the feature set.

In summary, this chapter introduced the dataset used throughout this thesis and pre-processing steps to build the two types of feature sets to be applied in the following two computational models. While the window-based dataset will be applied in the SVM model in Chapter 4, the frame-based feature set will be applied in the SVM model in Chapter 5.

Table 3.2

Overview of extracted features and tools

Modality	Category	Features	Tools
Visual	head	Nods Head Shake	Manual annotation (Bodur, 2021)
	mouth	Smile with mouth open/closed Laugh	
	eyebrow	frown Raised	
	gaze	Looking at the screen	
Vocal	Frequency	F0semitone mean F0semitone mean Rising Slope F0semitone mean Falling Slope F1frequency mean F2frequency mean F3frequency mean	Noise reduction algorithms OpenSmile Toolkit
	Spectral flux	Spectral flux mean	
	Voice quality	jitterLocal_sma3nz_amean shimmerLocaldB_sma3nz_amean HNRdBACF_sma3nz_amean	
	loudness	loudness_sma3_amean loudness_sma3_meanRisingSlope loudness_sma3_meanFallingSlope Equivalent Sound Level_dBp	
	Temporal features	Voiced Segments Per Sec Mean Voiced Segment Length Sec MeanUnvoiced Segment Length	
Verbal	Word class Surprisal Information content	POS Surprisal = $-\log_2 P(\text{word}\mid\text{occurrence})$ De-contextualized / Contextualized	SpaCy HuggingFace

Chapter 4. Non-sequential model

4.1 Introduction

The main objective of this chapter is to use computational models to predict the listener's BC occurrence based on the speaker features occurring in the context window as Figure 4.1 shows. The task of predicting BC occurrences is elaborated below. Consider the onset of the target behavior L, and let S represent the speaker's feature set in the given context window. Then, the aim of the computational model is to learn a function F_{bc} mapping the speaker's features within a time series to the corresponding BC opportunity label BO (a binary label signifying the presence or absence of BC in the time L), i.e., Fbc (L) $\rightarrow$ BO. Once the model has achieved optimal performance in prediction accuracy after a period of training, the BC inviting cues can be further investigated via manipulating different combinations of model inputs. Given the dyadic nature of the dialogue, four models predicting were trained and tested with different input features respectively.

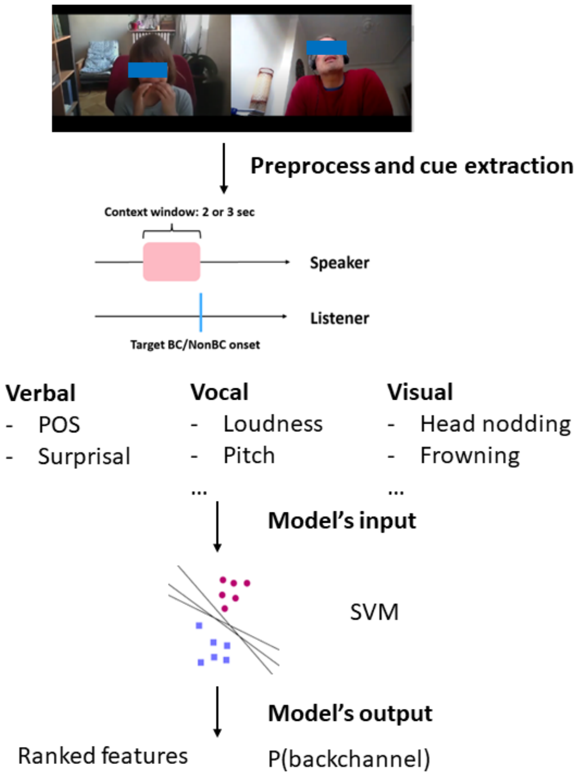


Figure 4.1 Schematic description of the modeling process, starting from video recordings of child-caregiver conversations

4.2 Model setting

In this chapter, the *Support Vector Machines* (hereafter SVM) was selected as the baseline model mainly due to its strength in processing highly dimensional features and no presupposition of feature independence (Dorman et al., 2013). Given a lot of features in the dataset are correlated (e.g. jitter and shimmer), SVM can thus facilitate interpretability of results in our dataset when using combined variables as input, especially compared with GLM models (Kiers and Smilde, 2007). Inspired by the strength, SVM has been extensively applied as one of the classifiers in combination with acoustic features to detect atypical speech (e.g.

van Bemmel et al., 2021) and BC response type (Jain et al., 2021). SVM typically uses a hyperplane to separate the data into classes. As Figure 4.2 shows, the standardized data is first projected into a feature space in which a hyperplane separates the classes as accurately as possible by maintaining the largest distance from each data sample. In this study, the window-based feature set was selected as the model input due to the constraint of the model structure. The equation below shows the final decision function of a binary SVM.

$$f(x) = \text{sign}(\sum_{k=1}^{n} y_k \alpha_k k(x \cdot x_k) + b)$$

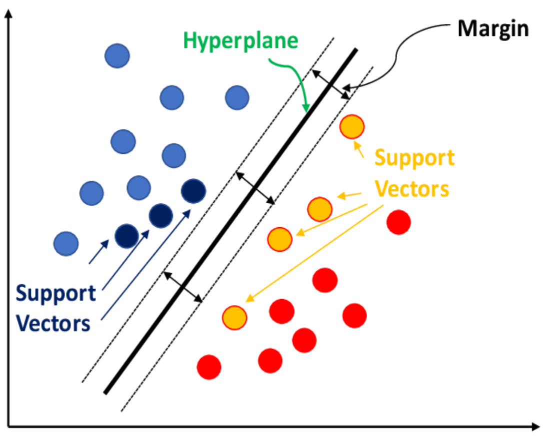


Figure 4.2 Visualization of Support Vector Machine Model

Given the high inter-locutor variability, *Leave-One-Subject-Out* (LOSO) cross validation was applied in this thesis through majority-voting to minimize the influence of interlocutor-specific features (Neto et al., 2019) and reduce identity-confounding (Shahin and Ahmed, 2019). As is shown in Figure 4.3 below, this validation strategy uses features from all speakers as the training data except for one and uses the left-out data as test data.

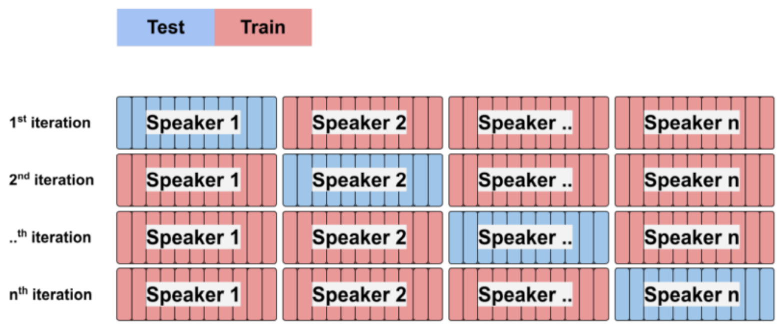


Figure 4.3 Visualization of LOSO validation

Then the optimal feature set yielding the best model performance was selected through Recursive Feature Elimination (RFE). Instead of a brute force search strategy, which would require 2^{56} possible combinations with the 56 features, RFE was selected for feature ranking with a combination of the SVM classifier (Kumar et al., 2014). As the pseudo code of implementation below shows, RFE removes features based on the coefficients until only the most relevant feature is left. Thus the first discarded feature is the least important while the last remaining feature is the most important one. The whole list of ranked features from different modalities are shown in Table 6 in appendix.

Algorithm: Recursive Feature Elimination to obtain a set of ranked features

Data: $D = \{X, L\}$ // dataset with n features where $X = \{f_1, f_2, f_3, \ldots, f_n\}$ and L are the labels
$R = \{\}$ // initially empty ranking list
X'// current feature sbset ($X' \subseteq X$ or $X' = \{\}$)
Fworst // least important feature
Result: $R = \{f1, f2, f3, \ldots fn\}$ // a feature ranking set
Begin
Initialize:
$X' = D$ //initialize with the entire feature subset
While ($X' \neq \{\}$) do
SVM (X',L) //train SVM with current feature subset
Fworst = min(SVM_{weights}) //get least important feature according to SVM weights
$R = R + f_{worst}$ // add worst feature to Ranking list
$X' = X' - f_{worst}$ // delete this worst performing feature from current feature subset
reverse(R) //Reverse the ranking list
Return R

4.3 Results

As exemplified in Figure 4.4, the model was recursively trained with different number of feature set in the ranked list and the combination that yielded the highest accuracy was summarized in Table 4.1 and 4.2. Notably, the feature combinations leading to the highest model performance were always the subset of the whole feature set, i.e. adding more features didn't necessarily improve model performance, rather, decreased the model performance.

As Table 4.1 shows, features extracted from shorter context windows(2s) were more predictive of both children and adults' BC responses, which echoed previous studies (Goswami et al., 2020). Therefore, model performance and combined features were analyzed in the 2s context window in the following analysis.

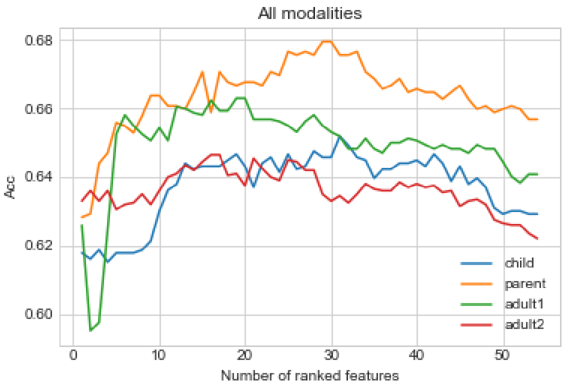


Figure 4.4 Example of feature combination selection

To minimize the confounding influence of other factors, children's ability of encoding BC inviting cues as the speaker was investigated through the comparison of models trained on caregivers in child-caregiver conversations and adult1 in adult-adult conversations; and their ability of providing BC responses as the listener was compared with adult2 in adult-adult conversations considering caregivers and adult1 were the same group of people in two conversations. As Table 4.1 shows, most models trained on interlocutors in child-caregiver conversations had slightly lower performance than adult-adult conversations across modalities. In particular, there was a larger difference in verbal modality between children and adults compared with other modality difference. A further comparison of contributing features within the verbal modality reflected that content words contributed more in eliciting BC responses than the functional words in child-caregiver conversations.

A comparison of the different single modalities within speakers indicated that both children and adults put highest weight on vocal modality and lowest on visual modality when encoding BC inviting cues. Besides, for the same interlocutor, the addition of other modalities improved the model performance.

For all the interlocutors, specific BC responses were more predictive than generic ones across modalities. In particular, there was a higher child-adult difference in predicting generic BC responses than specific ones.

4.4 Discussion

Research question I: Child-adult difference in overall model performance

Overall speaking, the slightly lower performance across modalities in child-caregiver conversations echoed our first hypothesis that children generally had a lower consistency in producing BC responses and inviting cues (at least harder to be captured by the SVM model). In particular, the larger difference in verbal modality between children and adults in both cases when children produced and responded to the BC inviting cues indicated that children were generally less sensitive to the verbal cues compared with adults. Specifically, the trend that adults' BC responses were more likely to be primed by the functional words suggested that they were more sensitive to grammatical completeness of the speaker's utterance.

However, there was no difference in modality preference as both children and adults put highest weight on vocal modality and lowest on visual modality when encoding BC inviting cues. These results were in line with previous findings on feature contribution in models trained on adult-adult conversations (Boudin et al., 2021; Goswami et al., 2020). Also, the fact that model performance benefited from additional information from other modalities suggested that both children and adults were capable of integrating multimodal cues in face-to-face conversations.

It is worth attention that such differences were marginal and might be caused by individual difference as indicated by the large overlapping of the confidence interval. What's more, features from verbal modality were only represented as the number word types occurring in the given time window, which failed to capture the sequential information of the verbal modality.

Research question II: BC type difference

For all the interlocutors, specific BC responses were more predictive than generic ones. In particular, a higher child-adult difference was found in predicting generic BC responses than specific ones, which provided some evidence for our second hypothesis that generic BC responses were not as highly correlated with context as specific ones. And the larger child-adult difference may indicate that generic BC responses were acquired later, perhaps, due to the highly demanding socio-cognitive ability.

In summary, this chapter used SVM models to inspect the potential BC inviting cues and the differences between children and adults. The slightly lower performance of models trained on children's conversational responses could suggest that middle-childhood children were still in the stage of developing communicative abilities. And the higher predictive effects of specific BC within speakers were aligned with their communicative functions, i.e., higher contingency with the conversation content is more predictive. What's more, both children and adults benefit from processing information from different channels.

Notably, it is still questionable whether the lower performance in verbal modality comes from the limitation of model structure that simply utilizes the averaged information instead of incorporate the dynamic change within the context window, i.e. sequential information. What's more, the overall model performance was not high enough, though comparable with previous studies (e.g. Jindal et al., 2020), which posed the question of whether the model result is the most appropriate given the task at hand. In the next chapter, I will partly address this question using a neural network model that takes into account sequential information.

Table 4.1

Comparison of model performance; Accuracy range comes from cross validation results tested on different interlocutors

Modality	Listener	All BC responses		Generic BC		Specific BC	
		Acc (3s)	**Acc (2s)**	**Acc (3s)**	**Acc (2s)**	**Acc (3s)**	**Acc (2s)**
visual	Child	0.553 [0.500, 0.633]	0.532 [0.485, 0.640]	0.256 [0, 0.600]	0.281 [0, 0.600]	0.534 [0.125, 0.621]	0.538 [0.125, 0.586]
	Caregiver	0.560 [0.518, 0.650]	0.571 [0.500, 0.625]	0.385 [0.147, 0.875]	0.422 [0.147, 1.000]	0.581 [0.467, 0.727]	0.538 [0.125, 0.586]
	Adult1	0.561 [0.500, 0.605]	0.552 [0.500, 0.650]	0.498 [0.305, 0.589]	0.455 [0.089, 0.931]	0.594 [0.400, 0.739]	0.595 [0.436, 0.696]
	Adult2	0.553 [0.485, 0.616]	0.536 [0.475, 0.620]	0.451 [0.214, 0.706]	0.396 [0.202, 0.578]	0.584 [0.227, 0.653]	0.578 [0.182, 0.652]
vocal	Child	0.629 [0.488, 0.819]	0.656 [0.529, 0.738]	0.566 [0.333, 0.714]	0.567 [0.331, 0.714]	0.654 [0.345, 0.875]	0.680 [0.414, 0.789]
	Caregiver	0.670 [0.524, 0.813]	0.658 [0.549, 0.796]	0.534 [0.176, 0.750]	0.578 [0.542, 0.778]	0.654 [0.345, 0.875]	0.680 [0.414, 0.789]
	Adult1	0.639 [0.582, 0.769]	0.679 [0.593, 0.767]	0.534 [0.176, 0.750]	0.654 [0.506, 0.856]	0.718 [0.604, 0.853]	0.725 [0.500, 0.863
	Adult2	0.638 [0.447, 0.708]	0.689 [0.473, 0.767]	0.563 [0.412, 0.685]	0.595 [0.422, 0.852]	0.629 [0.441, 0.742]	0.672 [0.468, 0.843]

verbal	Child	0.554 [0.500, 0.691]	0.563 [0.500, 0.704]	0.363 [0, 0.556]	0. 322 [0, 0.605]	0.567 [0.000, 0.792]	0.576 [0.000, 0.725
	Caregiver	0.559 [0.500, 0.728]	0.557 [0.500, 0.625]	0.613 [0.483, 0.844]	0.623 [0.379, 0.778]	0.567 [0.000, 0.792]	0.576 [0.000, 0.725
	Adult1	0.596 [0.500, 0.663]	0.611 [0.500, 0.652]	0.330 [0.000, 0.500]	0.332 [0.000, 0.546]	0.598 [0.000, 0.789]	0.608 [0.000, 0.737]
	Adult2	0.595 [0.500, 0.679]	0.626 [0.500, 0.708]	0.589 [0, 0.815]	0.622 [0, 0.796]	0.617 [0.500, 0.820]	0.645 [0.500, 0.843]
verbal + vocal	Child	0.631 [0.500, 0.833]	0.645 [0.496, 0.772]	0.631 [0.500, 0.833]	0.566 [0.496, 0.772]	0.653 [0.172, 0.833]	0.680 [0.414, 0.789]
	Caregiver	0.672 [0.524, 0.805]	0.664 [0.524, 0.796]	0.624 [0.470,0.889]	0.654 [0.506, 0.856]	0.653 [0.172, 0.833]	0.680 [0.414, 0.789]
	Adult1	0.637 [0.545, 0.775]	0.678 [0.600, 0.731]	0.534 [0.176,0.750]	0.548 [0.147, 0.815]	0.642 [0.509, 0.870]	0.673 [0.364, 0.742]
	Adult2	0.648 [0.504, 0.728]	0.690 [0.523, 0.750]	0.589 [0.412, 0.741]	0.586 [0.353, 0.759]	0.648 [0.516, 0.771]	0.690 [0.538, 0.809]
verbal + visual	Child	0.578 [0.476, 0.676]	0.566 [0.452, 0.664]	0.578 [0.476, 0.676]	0.566 [0.452, 0.664]	0.576 [0.103, 0.800]	0.574 [0.000, 0.744]
	Caregiver	0.578 [0.508, 0.707]	0.605 [0.517, 0.717]	0.398 [0.111,0.625]	0.463 [0.118, 1.000]	0.623 [0.333, 0.842]	0.644 [0.473, 0.870]

	Adult1	0.623 [0.551, 0.687]	0.606 [0.494, 0.696]	0.624 [0.448, 0.800]	0.623 [0.379, 0.778]	0.646 [0.455, 0.870]	0.640 [0.623, 0.842]
	Adult2	0.585 [0.520, 0.667]	0.625 [0.500, 0.689]	0.589 [0, 0.815]	0.622 [0, 0.796]	0.633 [0.510, 0.876]	0.625 [0.532, 0.831]
All	Child	0.641 [0.500, 0.806]	0.652 [0.531, 0.733]	0.353 [0, 0.695]	0.356 [0, 0.581]	0.652 [0.345, 0.875]	0.656 [0.172, 0.844]
	Caregiver	0.681 [0.589, 0.789]	0.666 [0.537, 0.778]	0.632 [0.425, 0.844]	0.650 [0.568, 0.889]	0.718 [0.604, 0.853]	0.729 [0.467, 0.850]
	Adult1	0.640 [0.572, 0.750]	0.672 [0.591, 0.725]	0.636 [0.581,1.000]	0. 654 [0.506, 0.856]	0.688 [0.564, 0.870]	0.718 [0.636, 0.870]
	Adult2	0.636 [0.535, 0.704]	0.673 [0.544, 0.778]	0.600 [0.490, 0.722]	0.623 [0.206, 0.778]	0.655 [0.517, 0.775]	0.692 [0.573, 0.843]

Table 4.2

Feature combination that yields the best model performance

Modality	**Speaker**	**All BC responses**	**Generic BC**	**Specific BC**
visual	Child	Nod; Gaze; Posture; Smile; Laugh; Head shake	Gaze; Posture; Eyebrow raise; Laugh; Smile; Frown; Head shake; Nod	Nod; Smile
	Caregiver	Smile; Eyebrow raise; Head shake; Nod	Gaze, head shake, posture, eyebrow raise; Smile	Gaze, Eyebrow raise, Head Shake, Laugh

	Adult1	Head shake; Laugh; Gaze; Eyebrow raise; Nod; Smile; Frown	Gaze, Smile, Frown, Posture, head shake, Nod	Nod', Smile, Eyebrow raise, Head shake, Frown, Posture, Gaze
	Adult2	Gaze	Smile, Gaze, Laugh, Posture, Nod, Eyebrow raise	Eyebrow raise; Nod; Laugh
vocal	Child	Loudness(variation); Articulation rate; Falling intonation	Pause length; Pitch variation	Loudness variation, Articulation rate,Pitch variation
	Caregiver	Loudness(variation); Articulation rate; Falling intonation	Loudness variation, Voice quality, pitch	Articulation rate, Pitch variation
	Adult1	Pitch variation; Articulation rate; (low) pitch; Silence; Loudness(variation); Voice quality	Articulation rate	Loudness, Pitch, Low pitch region, Pause, voice quality
	Adult2	Pitch variation; Loudness(variation); Rising intonation; Low pitch region; articulation rate	Pause length	Pitch variation; Loudness(variation); Rising intonation; Low pitch region; Articulation rate
verbal	Child	Noun; Verb	Noun; Verb	Surprisal, Adv
	Caregiver	Noun; Verb; Adj	Surprisal, Adj, Aux	Noun, Adv
	Adult1	Noun; Adv; det; intj; Verb; Conj; Num; Surprisal; sconj; Adj; PropN; Aux	Verb, Adv, Det	Noun, Adv, Verb, Adj
	Adult2	Intj; Pron; Surprisal; Noun; Adv; Det	Intj; Pron; Surprisal; Noun; Adv; Det	Surprisal
verbal + vocal	Child	Pitch(variation); Det; Articulation rate; (Rising) loudness; Voice quality	Pause length, Pitch variation	Loudness variation, Articulation rate, Pitch(variation)
	Caregiver	(low) pitch; articulation rate; (rising) loudness; Verb; Adj	loudness (variation), pitch, Surprisal, Det, falling intonation	articulation rate, pitch variation

	Adult1	(rising) loudness; Adv; Intj; Pron; articulation rate; pitch variation; voice quality; Noun	Articulation rate	loudness, pitch, low pitch region, pause, voice quality, verb, pron
	Adult2	(Rising) loudness; Det; (low/falling) pitch; Conj; Surprisal; Falling loudness	Articulation rate, Noun, conj., Adv, Pitch	Articulation rate; Noun, Pitch(variation); voice quality
verbal + visual	Child	Noun; Nod; Frown; Gaze; Eyebrow raise; Conj; Adv; Laugh; Det; Head shake; Aux; Posture; Intj; Smile	Gaze, Eyebrow raise, Laugh, Smile, Frown, Head shake, Nod, Det, Posture, Aux, Sconj, Laugh	Surprisal; Nod
	Child	Smile; Verb; Adj; Nod	Head shake, Smile, Surprisal, Smile, Aux, Gaze, Sconj, Eyebrow raise, Posture, Adj, PropN	Gaze, Frown, Det, Pron, Eyebrow raise
	Adult2	Adp; Pron; Cconj	Verb, Adv, Det, Adp	Nod, Smile, Det, Verb, Head shake, Adj, PropN, Adv, Noun, Frown
	Adult2	Noun; Intj; Conj, Laugh; Pron	Surprisal	Surprisal, Laugh, Noun, Nod, Sconj, Adj, Frown, Adv, PropN, Part, Verb
All	Child	Loudness(variation); Low pitch region; Smile; Frown; POS; Voice quality; Articulation rate	Pause length, Pitch variation	(falling) loudness, Smile, Pitch variation, Articulation rate
	Child	Loudness; (low) pitch; articulation rate; rising loudness	Head shake, Loudness(variation) , posture, pitch (variation), voice quality, Adj, Rising/falling intonation, Gaze, Det, Surprisal	Articulation rate, Gaze, pitch variation, Smile, low pitch region
	Adult2	articulation rate; smile; rising loudness; silence; low pitch region; (falling) pitch; POS; voice quality; gaze; eyebrow raise; surprisal	Articulation rate	Loudness, Nod, Pitch (variation), low pitch region, Pause, Smile, voice quality, Verb, Pron, Gaze, Posture
	Adult2	Silence; Rising intonation; Smile; eyebrow; Nod; POS; Voice quality; Gaze; Falling loudness; Frown; Posture; Laugh; Head shake; Surprisal	Noun, (Rising) Loudness, Smile, Adv., Adj	Articulation rate, Noun, Pitch variation, Voice quality, Surprisal, Gaze, Falling intonation

Chapter 5. Sequential model

5.1 Introduction

Last chapter proposed that the relatively lower model performance, especially in verbal modality, may stem from the lack of dynamic change information within the context window. Therefore, long short-term memory architecture (hereafter LSTM) was selected in this chapter due to its strength in capturing the dynamics of a sequence of input frames in the context window. In the LSTM, the predicted state of the current frame, i.e. whether the listener will give a BC response or not, not only depends on the corresponding speaker features, but also on the state of the previous frame, thus converting the average-feature-based prediction in Chapter 4 into frame-by-frame and sequence-dependent prediction. Specifically, at given time t in the conversation, output for the state S_t is calculated based on the output from the previous state S_{t-1} and current speaker feature input X_t as Figure 5.7 shows. This process continues forming an information loop for a given state concerning time. Compared with the simple recurrent neural networks (hereafter RNNs: Elman, 1990), which also applies to frame-by-frame dependency, LSTM is able to capture long-range frame dependencies, thus making it possible to incorporate features occurring at the beginning of the context window in the conversation. To be more specific, RNNs generally apply the chain rule to compute the gradients which carry information to update RNNs parameters. As a result, if any one of the gradients becomes infinitesimally small, all the gradients would exponentially rush to zero due to multiplying (the vanishing gradient problem: Hochreiter and Schmidhuber, 1997), thus leading to insignificant parameter updates and no real learning. As an extension of RNNs, LSTM applies a gated mechanism to capture long-range dependency (Olah, 2015) as Figure 5.1 shows. At time step t, the input gate i_t controls how much each unit is updated as shown in Figure 5.2, the output gate o_t controls the exposure of the internal memory state as shown in Figure 5.3, and the forget gate f_t controls the amount of which each unit of the memory cell is erased using a sigmoid function in Figure 5.4. The memory cell c_t keeps the useful history information which will be used for the next process as shown in Figure 5.5.

The strength of processing sequential information in LSTM model has been investigated in many previous literatures, such as predicting neural activity in human sentence reading (Frank, 2016; Qian et al., 2021), text generation (Pawade et al., 2018) and automatic speech recognition (Weninger et al., 2015).

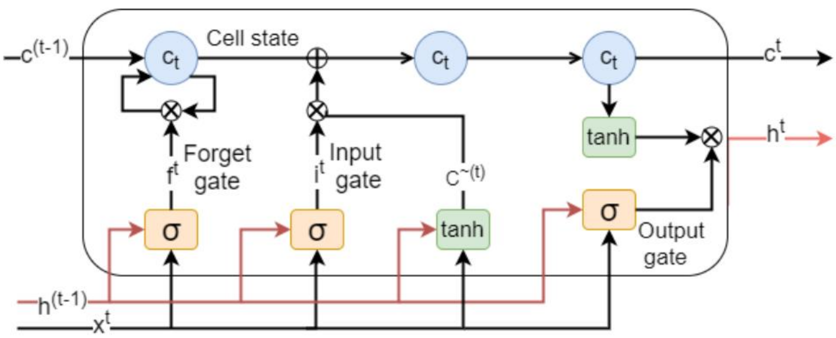
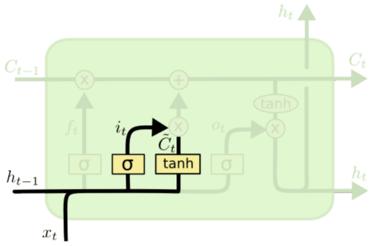
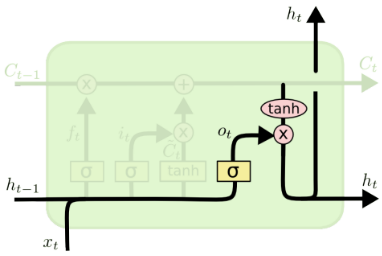


Figure 5.1 Long Short-Term Memory (LSTM) cell. Fundamental components of an LSTM cell are a forget gate, input gate, output gate and a cell state



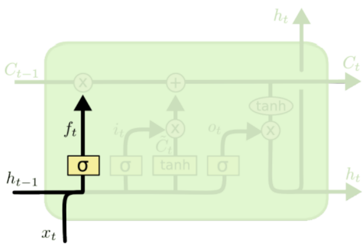
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Figure 5.2 Input Gate in LSTM



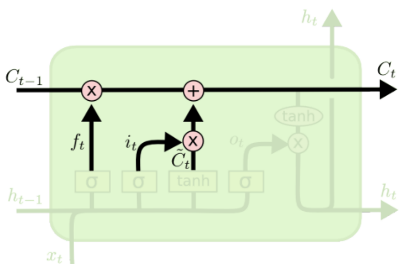
$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 5.3 Output Gate in LSTM



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

Figure 5.4 Forget Gate in LSTM



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 5.5 Cell information in LSTM

5.2 Model setting

The task of predicting BC occurrences is similar to the SVM model except that BC /nonBC onset was defined as the first three frames of the whole behavior. What's more, the frame-based feature set was applied as the model input. The task is re-elaborated below. Consider a time frame Lij, which starts at the ith second and ends at the jth, and let S(i-n)(i) represent the speaker's multimodal features occurring n frames prior to the target frame. Then, the aim of the computational model is to learn a function F_{bc} mapping the speaker's features within a time series to the corresponding BC opportunity label BO_{ij} (a binary label signifying the presence or absence of BCs in the time $\mathrm{L}ij$), i.e., $\mathrm{F}bc\ (\mathrm{L}ij) \rightarrow \mathrm{BO}ij$. Specifically, at each step t of frame-size 50ms, the network receives the last two seconds comprising 40 frames of features of the speaker as input and produces a binary output y_t (many-to-one mapping) as Figure 5.6 shows. The output layer of the network uses an element-wise sigmoid activation function to predict a probability score for the target interlocutor's BC behaviors at each future frame. Once the model has achieved optimal performance in prediction accuracy by tuning hyperparameters, the BC behavior inviting cues can be further investigated via manipulating different combinations of input cues.

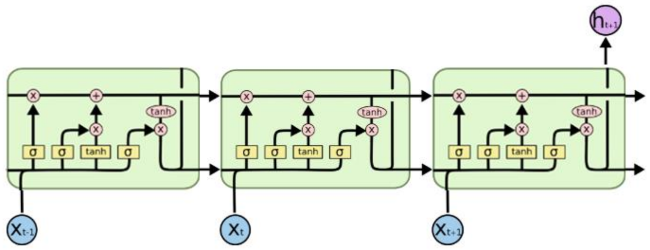


Figure 5.6 Visualization of many-to-one mapping

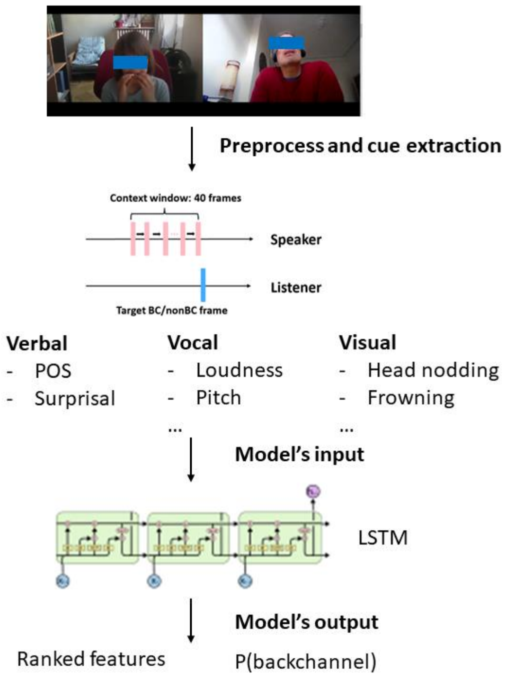


Figure 5.7 Schematic description of the LSTM modeling process, starting from video recordings of child-caregiver conversations

Similar to SVM model in Chapter 4, LSTM model also applied LOSO cross-validation to minimize the influence of interlocutor-specific features and reduce identity-confounding. As for the hyperparameter tuning, instead of using grid searches to check all the possible parameter combinations, Bayesian optimization for three hyperparameters (learning rate, dropout, and L2 regularization) were performed for each network configuration using *Optuna* (Akiba et al., 2019) considering the trade-off between the runtime and performance score. In order to limit the influence of parameter count changes between the different network configurations, the hidden Node count in a given network was limited to a sum of 50.

5.3 Results

To make the LSTM model results more comparable to the SVM model, SVM models were re-trained using the frame-based feature set. As Table 5 shows, the LSTM model performance was higher than the SVM model for all interlocutors, which indicated that sequential information was important in predicting BC behavior. Besides, there was a larger improvement in verbal-related modalities compared with others, thus echoed the hypothesis in Chapter 4 that the lower performance in the SVM model might stem from the lack of sequential information.

In contrast with the SVM models in Chapter 4, there was no systematic difference between child-caregiver and adult-adult conversations as Table 5 shows. In particular, a comparison of the different single modalities within speakers indicated that while children put highest weight on vocal modality, adults tended to put highest weight on verbal modality when encoding BC inviting cues. However, for the same interlocutor, the addition of features from other modalities didn't necessarily improve the model performance.

Similar to the SVM models, specific BC responses were more predictive than generic ones for all interlocutors. But there were no within-speaker predictive differences in terms of different BC types.

Table 5

Comparison of model performance; Accuracy range comes from the speaker variability

Modality	Listener	All BC responses		Generic BC LSTM	Specific BC LSTM
		SVM	LSTM		
visual	**Child**	0.587 [0.556, 0.742]	0.757 [0.556, 0.956]	0.679 [0.549, 0.828]	0.797 [0.545,0.903]
	Caregiver	0.532 [0.506, 0.913]	0.759 [0.597, 0.971]	0.758 [0.597,0.971]	0.764 [0.512,0.888]
	Adult1	0.536 [0.356, 0.914]	0.771 [0.664, 0.929]	0.744 [0.651, 0.929]	0.750 [0.600,0.920]
	Adult2	0.562 [0.356, 0.687]	0.738 [0.529, 0.899]	0.612 [0.471,0.862]	0.751 [0.356, 0.914]
vocal	**Child**	0.554 [0.346, 0.750]	0.783 [0.615, 0.966]	0.689 [0.473, 0.767]	0.819 [0.488, 0.937]
	Caregiver	0.656 [0.615, 0.966]	0.734 [0.597, 0.971]	0.679 [0.549, 0.796]	0.766 [0.597,0.971]
	Adult1	0.689 [0.493, 0.911]	0.752 [0.664, 0.929]	0.744 [0.662, 0.905]	0.761 [0.473, 0.956]
	Adult2	0.554 [0.491, 0.686]	0.715 [0.493, 0.914]	0.684 [0.478, 0.758]	0.741 [0.664, 0.929]
verbal	**Child**	0.591 [0.308, 0.786]	0.744 [0.583, 0.860]	0.689 [0.583, 0.860]	0.767 [0.593,0.868]
	Caregiver	0.563 [0.583, 0.860]	0.806 [0.597, 0.971]	0.741 [0.6, 0.971]	0.815 [0.573, 0.849]

	Adult1	0.626 [0.493, 0.914]	0.770 [0.664, 0.929]	0.684 [0.491, 0.778]	0.765 [0.664, 0.929]
	Adult2	0.540 [0.494, 0.642]	0.707 [0.493, 0.914]	0.687 [0.412, 0.92]	0.740 [0.545,0.912]
verbal + vocal	**Child**	0.587 [0.385, 0.716]	0.781 [0.611, 0.966]	0.647 [0.512, 0.764]	0.792 [0.603, 0.943]
	Caregiver	0.645 [0.611, 0.966]	0.744 [0.662, 0.905]	0.711 [0.519, 0.729]	0.752 [0.664, 0.929]
	Adult1	0.690 [0.493, 0.931]	0.719 [0.611, 0.966]	0.715 [0.493, 0.914]	0.769 [0.582, 0.933]
	Adult2	0.562 [0.493, 0.775]	0.740 [0.493, 0.931]	0.677 [0.537, 0.833]	0.789 [0.537, 0.925]
verbal + visual	**Child**	0.587 [0.385, 0.716]	0.785 [0.641, 0.933]	0.782 [0.607, 0.971]	0.796 [0.605, 0.966]
	Caregiver	0.566 [0.503, 0.942]	0.773 [0.597, 0.971]	0.723 [0.600, 0.951]	0.806 [0.547, 0.924]
	Adult1	0.625 [0.496, 0.920]	0.755 [0.664, 0.929]	0.711 [0.519, 0.729]	0.792 [0.512,0.837]
	Adult2	0.562 [0.493, 0.931]	0.747 [0.501, 0.930]	0.728 [0.502, 0.908]	0.775 [0.555, 0.932]
All	**Child**	0.606 [0.521, 0.700]	0.788 [0.613, 0.966]	0.684 [0.491,0.777]	0.809 [0.609, 0.966]
	Caregiver	0.652 [0.609, 0.966]	0.782 [0.597, 0.971]	0.762 [0.600, 0.971]	0.805 [0.524, 0.981]
	Adult1	0.673 [0.493, 0.908]	0.765 [0.654, 0.929]	0.678 [0.591, 0.788]	0.797 [0.537,0.934]
	Adult2	0.595 [0.750, 0.908]	0.720 [0.493, 0.908]	0.708 [0.493, 0.908]	0.765 [0.559, 0.911]

5.4 Discussion

This chapter reflects that the LSTM model can predict human BC responses more accurately compared with the SVM model, especially thus more reliable in interpreting humans' responses to inviting cues.

Research question I: Child-adult difference in overall model performance

Overall speaking, the comparable model performance across modalities in two types of conversations contradicted the results in Chapter 4 and our first hypothesis of children's lower consistency in producing BC responses and inviting cues, especially in terms of the largely overlapping range. Though there was no significant difference in model performance, child-adult difference lied in their different weights on modalities, which was in contrast with the results in Chapter 4. Interestingly, while children's highest weight on vocal modality corresponded to findings in Chapter 4 and previous studies (Boudin et al., 2021; Goswami et al., 2020), adults' highest weight on verbal modality indicated that LSTM captured the grammatical structure based on the sequence of POS. What's more, the fact that concatenating additional modalities didn't necessarily improve the performance on verbal modality suggested that there were redundant features in other modalities.

Notably, although LSTM models have exhibited higher performance than SVM models trained in the same setting, there's no feature selection procedure in the LSTM, thus making the model performance still not optimal. What's more, children and adults' different weights in different modalities and the cofounding influence from the additional modality information may also be caused by the lack of feature selection procedure, thus incorporating some confounding factors.

Research question II: BC type difference

For all the interlocutors, specific BC responses were more predictive than generic ones. In particular, a higher child-adult difference was found in predicting generic BC responses than specific ones, which provided some evidence for our second hypothesis that generic BC responses were not as highly correlated with context as specific ones.

In conclusion, this chapter reflects that the sequential models can better predict human performance. In contrast with the SVM model in the last chapter, children's BC responses were not less predictive compared with adults. What's more, the higher predictive effects of specific BC within speakers were aligned with their communicative functions, i.e., higher contingency with the conversation content is more predictive, which was in line with the SVM model results and our hypothesis. More discussions on the results will be offered in Chapter 6.

Chapter 6. Conclusion

6.1 Thesis overview

This thesis aimed to investigate middle-childhood children's backchannel behaviors in conversations through a corpus analysis on child-caregiver interactions. Complementary to previous work applying frequency-based comparison on children-adults' BC usage (Hess and Johnson, 1988; Bodur et al., 2022), we examined BC inviting cue differences by analyzing the speaker behaviors in a given context window and exploring their potential causal relationship using more complex models. Considering the complexity of face-to-face conversation, the SVM and LSTM models were selected due to the following strengths: capacity to exploit more ecologically valid multimodal data and higher model performance (Dupoux, 2018).

Although simplification (Eberlein, 1989) and appropriate abstraction from raw signals have been an indispensable step to achieve psychological adequacy, i.e. capturing key details of human behavior, while providing an understandable account of how the model works, (McClelland & Elman, 1986), we claim that an exception has to be made to model children's BC behaviors due to the following two reasons. On the one hand, humans' representations of linguistic categories are still under debate (e.g. phonological category: Feldman et al., 2021; grammatical structure: Ding et al., 2016), even for middle-childhood children who are assumed to have acquired linguistic structures (Cekaite, 2012). For instance, recent human speech perception models borrowing ASR techniques to use the spectrum as the direct input (Magnuson et al., 2020; ten Bosch, 2018) have reached high performance in simulating human response. On the other hand, face-to-face conversations generally involve exquisitely complex signals and the complicated interactions among these modalities, which made the discrete and abstract features implausible in this respect.

What's more, current study aims to interpret listener's feedback on the computational level (1981, Marr) by abstracting away from considerations about processing or neural implementation, suggesting that these models are selected on the basis of similar performance as children rather than the simulation of children's brain processes. The computational-level modeling is essential to interpret cognitive mechanisms as verbal reasoning and toy models are spectacularly inclined to lead to incorrect predictions in the condition of combination of contradictory tendencies (Dupoux, 2018). In this respect, the relatively higher model performance offers a more uniform and quantitative standard compared with arbitrary or aesthetic criteria of model selection in previous research (e.g. select seemingly related model structure and test with simplified condition). Therefore, we moved from SVM model in Chapter 4 to LSTM model in Chapter 5 to interpret children and adults' responses due to the LSTM's comparatively higher model performance and the integration of sequential information.

Admittedly, the underlying cognitive mechanisms are still unclear by these models given that it is difficult to apply to the implementational level due to the constraint of computational power. As current supercomputers can only simulate a fraction of a brain and several orders of magnitude slower than real time (Kunkel et al., 2014), which is still massively underpowered compared to a child's brain. This makes claims of biological plausibility difficult to make. But we believe that it is an essential first step to investigate the developmental patterns of inviting cues to elicit BC responses on the computational level as these models offer a balance between the quantitative inferences and more ecologically valid data. What's more, the quantitative inference doesn't mean that models applied in this thesis have turned into "artificial interlocutors", on which researchers are able to simulate highly controlled in-lab experiments by manipulating inputs, model structures or even some parameters. This is due to the comparatively low model performance (though already higher than most published BC prediction models) and dyadic nature of dialogue interactions. For instance, caregivers tend to scaffold children through multi-level linguistic alignment (lexical: Fernandez and Grimm, 2014; conceptual: Misiek et al., 2020; syntactic: Dale and Spivey, 2006). That is, parents' speech depends fundamentally on children's own speech by borrowing children's own utterance's syntactic structure (e.g., by re-using their verbs or function words) or simply repeating parts of utterances to facilitate children's language processing. The adaptation mechanism perplexed the inspection of children's responses to BC inviting cues considering that caregivers might adapt their communicative strategies to different addresses in different age groups, which made the child-adult comparison more complex even though they respond to same group of speakers.

6.2 Framing the findings

6.2.1 Model selection

A comparison of different types of models suggested that the sequential models were more capable of simulating human language processing, which echoed previous claims on LSTM model's strength on capturing sequential information (Graves, 2013). However, the differences between these two models not only lie in the sequential information, but also the different model perplexity. While the SVM used in this thesis assumes that BC inviting cues are linearly separable by applying a linear kernel, the LSTM consists of different layers (i.e., the input layer, output layer and several hidden layers in between) and the information was passed through non-linear activation actions. Therefore, it is still unclear whether the model performance difference is caused by the sequential information or different complexity of model structure. Future studies should test other feedforward artificial neural networks (ANN) without integrating sequential information like multilayer perceptron (MLP).

6.2.2 Child-Adult difference

The comparison of model performance showed that middle-childhood children's ability to emit and respond to BC inviting cues was not less consistent than adults. The comparable model performance may be related with the nature of the task that explicitly required the listener's specific BC responses to proceed the game. Also, caregivers are assumed to scaffold children to facilitate children's language processing through multi-level linguistic alignment (lexical: Fernandez and Grimm, 2014; conceptual: Misiek et al., 2020; syntactic: Dale and Spivey, 2006). That is, caregivers' speech depends fundamentally on children's own speech by borrowing children's own utterance's syntactic structure (e.g., by re-using their verbs or function words) or simply repeating parts of utterances. As a result, middle-childhood children were likely to exhibit comparable predictability.

Notably, the comparable model performance doesn't necessarily mean that they have achieved adults' mastery given their selective attention to different modalities or even different features within the same modality. In particular, children were found to be more responsive to dynamic changes in vocal modality. The higher contribution was likely to be caused by caregivers' adaptation of speech production. As child-directed speech was characterized with clearer pitch, more pauses and exaggerated intonation (Peccei,1999), caregivers were found to spontaneously adapt their verbal input in ways that can facilitate children's affect, attention, and language development when speaking to young children (Saint-Georges et al., 2013). What's more, adults were found to put higher weight on functional words whereas children on content words, that is, adults tend to interpret the grammatical completeness as BC inviting cues in alignment with previous studies (e.g. Ward and Tsukahara, 2000). Thus, it is assumed that the different weight is closely related to children's ever-developing pragmatic awareness, that is, the metalinguistic awareness to detect inconsistencies between and within sentences (Betti, 2021; Igaab, 2010; Tuner et al., 1988). Previous studies have shown that metalinguistic awareness emerges during middle childhood, during which they develop the ability to reflect on structural characteristics of language as a parallel with the development of the ability to control their own cognitive functioning (Edwards & Kirkpatrick, 1999).

6.2.3 BC type

We also found that children's specific BC responses were more predictable than the generic ones. This may be related with the function of two types of BC: while generic BC indicates comprehension and attention to sustain the conversation flow without responding to the narrative content of the moment (Schegloff, 1982; Goodwin, 1986; Stivers, 2008), specific BC is closely related to the speaker context, thus more predictable by the preceding speaker features. In this regard, it may also be related with different underlying cognitive abilities in producing generic and specific BCs. While the former surpasses a simple reaction to the content towards the adequacy and quality of current content, thus requiring more advanced meta-linguistic abilities, the latter is more closely related with the narrative content. Therefore, there may be a delay in children's acquisition of generic BC compared to specific one.

6.2.4 Information entropy and BC occurrence

Additionally, the entropy-related features(word surprisal) were found to contribute to both children and adults' BC responses. This has raised the possibility to integrate information theory in conversation analysis to make more explicit reasoning. Recent studies on information theory can be a promising direction for future conversational analysis as it offers a quantitative and interpretable framework, which also shows the potential to uniform the rational principles with other levels of language processing. These studies have found that the rational strategy of information transmission (ERC and UID principles: Genzel and Charniak, 2002; Aylett and Turk, 2004; Jaeger and Levy, 2007) also holds in conversations, especially within the contextually contingent topic units (Giulianelli and Fernández, 2021; Giulianelli et al., 2021). Xu and Reitter (2018) further argued that the grounding process can be depicted as the converging trend of information density between different roles of speakers, in which the topic initiator's entropy kept decreasing and the topic receiver's kept increasing. In our preliminary analysis, constant information transmission rate and converging trend of information were also found to be applicable for our dataset (see Section 2 in Appendix). However, in the subsequent multimodal analysis, additional information from vocal and visual modalities also had an influence on model performance, thus indicating that interlocutors may interpret information from multiple channels. Therefore, a further investigation of how one's prediction of a certain word is influenced by information from other modalities is necessary to expand the information theoretic framework into conversations.

6.3 Limitations and future work

6.3.1 Data collection

Although more than one normalization strategy has been applied to reduce the speaker-specific variability like z-normalization on input features and LOSO cross validation during model test phase, we still observed a high variability when the model was tested on different interlocutors. which may be caused by the relatively small sample size (10 child-caregiver and 10 adult-adult dialogues) and higher interspeaker variability (one or two children at each age group). Therefore, larger sample size per age group is necessary for testing whether the inviting cues differ depending on the age of children.

What's more, children acquire the art of providing proper conversational feedback in culturally-laden interactions with their parents, peers and teachers. A plethora of studies indicate that BC occurrence is heavily culturally and contextually specific (Cutrone, 2005; McCarthy, 2002; Stubbe, 1998). However, only few studies focus on across cultural, sociodemographic and economic contexts factors on the development of BC utterances in young children (Curenton, 2010; Jindal et al., 2020). For instance, Jindal et al. (2020) found that the highest educational level of caregivers, gender of both interlocutors and household income influence BC occurrence to a large extent. Therefore, future analysis can integrate these features in a larger dataset.

Another limitation is zoom-recordings' influence on speech characteristics. Recent acoustic analyses of zoom speech data (Zhang et al., 2021) indicated that formant tracking presented some issues as reflected in F1 and F2 values distortion. The intensity drops observed in the Zoom recordings could also pose serious issues for speech analysis. Therefore, more effective pre-processing steps on speech data are needed to reduce the potential influence of recording devices.

6.3.2 Model configurations

In Section 5.3, a single LSTM was constructed with all input features being processed at the same rate. Given the acoustic features generally perform at the sub-word prosodic level, the concatenation of linguistic features on the word level led to either averaging the acoustic features at a coarse temporal granularity or upsampling the linguistic features at the acoustic temporal resolution. This thesis applied an upsampling strategy to assign the linguistic feature on multiple frames, which may cause problems of dealing with longer term dependencies due to the duplicated features over different frames. This problem can be addressed through a multiscale architecture in which different modalities can be modeled in separate sub-network LSTMs with independent timescales and then fused in a master LSTM (Roddy et al., 2018).

Considering the class imbalance in the dataset, I down-sampled the nonBC frames to get the more balanced proportion. However, this reduced the total dataset size and the randomly sampled nonBC frames may cause some problems. Another possible solution is to train the model with the whole conversation frames and evaluate the model using the evaluation task as in Roddy et al.(2018). Noticeably, this requires a more sufficient evaluation task. In our previous try-outs, the model was trained on the whole dialogues and further evaluated on the randomly sampled BC and nonBC frames with balanced proportion. The accuracy was pretty low. Therefore, future studies can utilize some data augmentation techniques like SVM-SMOTE (Chawla et al., 2011; Jain et al., 2021) to generate more data and feature warping (Murray et al., 2021) which have been widely used in speech recognition (Park et al., 2019; Toth et al., 2018).

Given the nature of our study, interpretability of our machine learning models is as important as their predictive accuracy. In chapter 4, I applied RFE in the SVM model. However, no feature interpretation algorithm has been conducted on the LSTM model yet due to presupposition of feature independence in some prevalent feature ranking method (Lundberg et al., 2017). Therefore, a further adaptation of feature ranking algorithms like Mean Decrease in Impurity (MDI) (Breiman, 2001) and Mean Decrease in Accuracy (MDA) (Breiman, 2002) are needed to interpret input cues in BC prediction models.

In conclusion, although children between the ages of 6 and 12 years are still in the stage of developing socio-cognitive competencies, their performance of producing and responding to BC inviting cues are strikingly close to adult-level mastery in the interactions with caregivers, possibly due to the richness of multimodal cues and the linguistic alignment from caregivers. In communication, they seem to understand the coordinative responsibilities they have both as a speaker and a listener. The discovery of all these aspects might prove to be important for predicting developmental patterns of children's interpretation and production of multimodal cues and building humanoid child-centered conversational agents.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Al Moubayed, S., Baklouti, M., Chetouani, M., Dutoit, T., Mahdhaoui, A., Martin, J. C., ... and Yilmaz, M. (2009, May). Generating robot/agent BCs during a storytelling experiment. In *2009 IEEE International Conference on Robotics and Automation* (pp. 3749-3754). IEEE.

Aylett, M., and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1), 31-56.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. *In 2016 IEEE Winter Conference on Applications of Computer Vision* (WACV) (pp. 1-10). IEEE.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive science*, 27(2), 195-225.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6), 941.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the acoustical society of America*, 113(2), 1001-1024.

Bemmel, L. V., Harmsen, W., Cucchiarini, C., & Strik, H. (2021, September). Automatic Selection of the Most Characterizing Features for Detecting COPD in Speech. In *International Conference on Speech and Computer* (pp. 737-748). Springer, Cham.

Betti, M. J. (2021). Metalinguistics and Metalinguistic Awareness.

Bigi, B., & Meunier, C. (2018). Automatic segmentation of spontaneous speech. *Revista de Estudos da Linguagem*, 26(4).

Blache, P., Bertrand, R. and Ferré, G. (2008) Creating and exploiting multimodal annotated corpora. *Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC)* 2008 [online]. pp.110-115. Available at: http://www.lrec-conf.org/proceedings/lrec2008/pdf/132_paper.pdf [Accessed 16 December 2008].

Bodur, K., Nikolaus, M., Fourtassi, A., and Prévot, L. (2022). BC Behavior in Child-caregiver Zoom-mediated Conversations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Bodur, K. (2021). Multimodal Language Use in Middle Childhood [Master's thesis, AIX – MARSEILLE UNIVERSITY]. Marseille.

Bonin, P., M´eot, A., Bugaiska, A.: Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behavior research methods* 50(6), 2366–2387 (2018)

Boudin, A., Bertrand, R., Rauzy, S., Ochs, M., & Blache, P. (2021, September). A multimodal model for predicting conversational feedbacks. In *International Conference on Text, Speech, and Dialogue* (pp. 537-549). Springer, Cham.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA.

Brigitte Bigi and Christine Meunier (2018). Automatic speech segmentation of spontaneous speech. In Revista de Estudos da Linguagem. *International Thematic Issue: Speech Segmentation*. Editors: Tommaso Raso, Heliana Mello, Plinio Barbosa, vol. 26, no 4, e-ISSN 2237-2083.

Brunner, L. J. (1979). Smiles can be back channels. *Journal of personality and social psychology*, 37(5), 728.

Bunt, H. (1994). Context and dialogue control. Think Quarterly, 3(1), 19-31.

Carlson, R., Gustafson, K., & Strangert, E. (2006). Modelling hesitation for synthesis of spontaneous speech. Proc. *Speech Prosody 2006*.

Cathcart, N., Carletta, J., & Klein, E. (2003, April). A shallow model of backchannel continuers in spoken dialogue. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (pp. 51-58).

Cekaite, A. (2012). Child Pragmatic Development. *In The Encyclopedia of Applied Linguistics*, C.A. Chapelle (Ed.). https://doi.org/10.1002/9781405198431.wbeal0127

Cerrato, L. (2002) A comparison between feedback strategies in Human-to[1]Human and Human-Machine communication. *Proceedings of International Conference of Speech and Language Processing (ICSLP)* Denver, Colorado. pp.557-560

Chaibub Neto, E., Pratap, A., Perumal, T. M., Tummalacherla, M., Snyder, P., Bot, B. M., ... and Omberg, L. (2019). *Detecting the impact of subject characteristics on machine learning-based diagnostic applications*. NPJ digital medicine, 2(1), 1-6.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication.

Clark, H.H. and E.F. Schaefer, Contributing to discourse. *Cognitive Science*, 1989. 13,: p. 259-294

Clark, H.H., Using Language. 1996, Cambridge: Cambridge University Press.

Clark, H.H. and Krych, M.A. (2004) Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1): pp.62-81.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1-39.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423

Curenton, S. M. (2010). Narratives as learning tools to promote school readiness.

Cutrone, P. (2005). A case study examining backchannels in conversations between Japanese–British dyads.

Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3), 391-430.

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, 59, 43–95.

Demberg, V., Sayeed, A., Gorinski, P., and Engonopoulos, N. (2012, July). Syntactic surprisal affects spoken word duration in conversational contexts. *In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 356-367).

Denny, R. (1985). Pragmatically marked and unmarked forms of speaking-turn exchange. *Interaction Structure and Strategy*, 135-174.

Dittman, A.T., and Llewellen, L.G. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 79-84.

Dittmann, A.T. (1972). Developmental Factors in Conversational Behavior. *Journal of Communication*, 22: 404-423. https://doi.org/10.1111/j.1460-2466.1972.tb00165.

Doyle, G., Yurovsky, D., and Frank, M. C. (2016). A robust framework for estimating linguistic alignment in twitter conversations. *In Proceedings of the 25th international conference on world wide web* (pp. 637–648). Montreal, Canada.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173, 43-59.

Eberlein, R. L. (1989). Simplification and understanding of models. *System Dynamics Review*, 5(1), 51-68.

Edwards, H.T., & A.G. Kirkpatrick (1999). Metalinguistic awareness in Children: A Developmental Progression. *Journal of Psycholinguistic Research*, 28, 4, 313-329.

Eyben, F., Wöllmer, M., and Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. *In Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459-1462).

Fernández, R., & Grimm, R. (2014). Quantifying categorical and conceptual convergence in child-adult dialogue. *In Proceedings of the annual meeting of the cognitive science society* (Vol. 36, No. 36).

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions.

Geertzen, J. (2015). Exploring age-related conversational interaction. *SEMDIAL* 2015 goDIAL, 42.

Genzel, D., and Charniak, E. (2002, July). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 199-206).

Genzel, D., and Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 65-72).

Giulianelli, M., and Fernández, R. (2021, November). Analyzing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 647-660).

Giulianelli, M., Sinclair, A., & Fernández, R. (2021, November). Is Information Density Uniform in Task-Oriented Dialogues?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8271-8283).

Goffman, E. (1967). Interaction ritual: Essays on face-to-face interaction.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11), 419-429.

Goodwin, C. (1986). Gestures as a resource for the organization of mutual orientation.

Goswami, M., Manuja, M., & Leekha, M. (2020). Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. *arXiv preprint arXiv:2007.11346.*

Gould, O. N., & Dixon, R. A. (1993). How I spent our vacation: collaborative storytelling by young and old adults. *Psychology and Aging*, 8(1), 10.

Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In Tenth Annual Conference of the International Speech Communication Association.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850.*

Grice, H. P. (1975). Logic and conversation. In Speech Acts

Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics on language technologies* (pp. 1–8). Stroudsburg, PA.

Hess, L. J., & Johnston, J. R. (1988). Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, 11(3), 319-335.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1), 23-62.

Jaeger, T., and Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Jain, V., Leekha, M., Shah, R. R., & Shukla, J. (2021, May). Exploring semi-supervised learning for predicting listener backchannels. *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Jindal, R., Leekha, M., Manuja, M., & Goswami, M. (2020). What makes a better companion? towards social & engaging peer learning. *In ECAI 2020* (pp. 482-489). IOS Press.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... and Shen, S. (2012). “Robovie, you'll have to go into the closet now”: Children's social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2), 303.

Kemper, S., FinterUrczyk, A., Ferrell, P., Harden, T., & Billington, C. (1998). Using elderspeak with older adults. *Discourse Processes*, 25(1), 55–73

Kendon, A. (1972) Some relationships between body motion and speech. In Seigman, A. and Pope, B. (Eds.) *Studies in Dyadic Communication.Elmsford, New York: Pergamon Press. pp.177-216.*

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. *SMOTE: Synthetic Minority Over-sampling Technique.* CoRR abs/1106.1813 (2011), 321—-357. arXiv:1106.1813 http://arxiv.org/abs/1106.1813

Kiers H. A. L. and Smilde A. K. 2007. A comparison of various methods for multivariate regression with highly collinear variables. *Stat. Methods Appl*. 16: 193–228.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech and Language*, 45, 326-347.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and BCs based on prosodic and syntactic features in Japanese map task

Kumar, V., & Minz, S. (2014). Feature selection: a literature review. *SmartCR*, 4(3), 211-229.

Kunkel, S., Schmidt, M., Eppler, J. M., Plesser, H. E., Masumoto, G., Igarashi, J., ... & Helias, M. (2014). Spiking network simulation code for petascale computers. *Frontiers in neuroinformatics*, 8, 78.

Lee, J. J., Breazeal, C., and DeSteno, D. (2017). Role of speaker cues in attention inference. *Frontiers in Robotics and AI*, 4, 47.

Levelt, W.J.M. (1989). Speaking: From Intention to Articulation. Cambridge, MA: MIT Press.

LevinsoLove, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.n, S. C. (1979). Activity types and language.

Liu, J. (2021). Polishing the input features for automatic Mandarin tone classification. Poster presented at Netherlands Graduate School of Linguistics (*LOT Summer School 2021), Leuven, Belgium.*

Liu, J., Strik, H. (2022). Native and non-native speakers’ idiom production: What can read speech tell us? *Oral presentation at MWE Workshop*, Marseille, France

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Marchetti, A., Manzi, F., Itakura, S., and Massaro, D. (2018). Theory of mind and humanoid robots from a lifespan perspective. *Zeitschrift für Psychologie* 226, 98–109. doi: 10.1027/2151-2604/a000326

Maynard, S.K., 1989. Japanese conversation. *Norwood*, NJ: Ablex

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *INTERSPEECH*.

McCarthy, M. (2002). Good listenership made plain. *Using corpora to explore linguistic variation*, 9, 49.

McClave, E.Z. (2000) Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32(7): pp.855-878.

Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006, March). A spatial model of engagement for a social robot. I*n 9th IEEE International Workshop on Advanced Motion Control*, 2006. (pp. 762-767). IEEE.

Mills, G., Groningen, C., & Redeker, G. (2017). Amplifying signals of misunderstanding improves coordination in dialogue. *FADLI* 2017, 52.

Misiek, T., Favre, B., & Fourtassi, A. (2020, November). Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 54-58).

Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems*, 20(1), 70-84.

Murray, M., Walker, N., Nanavati, A., Alves-Oliveira, P., Filippov, N., Sauppe, A., ... & Cakmak, M. (2022, January). Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. *In Conference on Robot Learning* (pp. 513-525). PMLR.

Nishimura, R., Kitaoka, N., and Nakagawa, S. (2007, September). A spoken dialog system for chat-like conversations considering response timing. *In International Conference on Text, Speech and Dialogue* (pp. 599-606). Springer, Berlin, Heidelberg.

O'Keeffe, A. and Adolphs, S. (2008) Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. In Schneider, K.P. and Barron, A. (Eds.) *Variational Pragmatics*. Amsterdam, Netherlands: John Benjamins. pp.69-98

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2), 283-328.

Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., and Itakura, S. (2013a). Infants understand the referential nature of human gaze but not robot gaze. J. *Exp. Child Psychol*. 116, 86–95. doi: 10.1016/j.jecp.2013.02.007

Olah, C. (2015). Understanding lstm networks.

Oreström, B. (1983) Turn-taking in English Conversation. Lund, Sweden: LiberFörlag Ltd.

Park, H. W., Gelsomini, M., Lee, J. J., and Breazeal, C. (2017, March). Telling stories to robots: The effect of BCing on a child's storytelling. In 2017 12th ACM/IEEE *International Conference on Human-Robot Interaction* (HRI (pp. 100-108). IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Pawade, D., Sakhapara, A., Jain, M., Jain, N., & Gada, K. (2018). Story scrambler-automatic text generation using word level RNN-LSTM. *International Journal of Information Technology and Computer Science (IJITCS)*, 10(6), 44-53.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2), 169-190.

Qian, T., and Jaeger, T. F. (2011). Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Roddy, M., Skantze, G., & Harte, N. (2018, October). Multimodal continuous turn-taking prediction using multiscale rnns. *In Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 186-190).

Ruede, R., Müller, M., Stüker, S., & Waibel, A. (2017, August). Enhancing Backchannel Prediction Using Word Embeddings. In *Interspeech* (pp. 879-883).

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10), e1008228.

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., ... & Cohen, D. (2013). Motherese in interaction: at the cross-road of emotion and cognition?(A systematic review). *PloS one*, 8(10), e78103.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71, 71-93.

Shahin, M., Zafar, U., & Ahmed, B. (2019). The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 400-412.

Solorio, T., Fuentes, O., Ward, N. G., & Al Bayyari, Y. (2006, September). Prosodic feature generation for back-channel prediction. *In INTERSPEECH* (Vol. 5, pp. 2398-2401).

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction*, 41(1), 31-57.

Stubbe, M. (1998). Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3), 257-289.

Sugito, M., 1994. Nihonjin no koe (The speech of the Japanese people). Tokyo: Izumi Shoin.

Sundberg Cerrato, L. (2007). Investigating communicative feedback phenomena across languages and modalities (*Doctoral dissertation*, *KTH*).

Takeuchi, M., Kitaoka, N., and Nakagawa, S. (2004). Timing detection for realtime dialog systems using prosodic and linguistic information. *In Speech Prosody* 2004, International Conference.

Temperley, D. and Gildea, D. (2015), Information Density and Syntactic Repetition. *Cogn Sci*, 39: 1802-1823. https://doi.org/10.1111/cogs.12215

Thomas Kisler, Uwe D. Reichel, and Florian Schiel (2017): Multilingual processing of speech via web services, *Computer Speech & Language*, Volume 45, September 2017, pages 326–347.

Tomasello, M. (1999). The cultural origins of human cognition. Harvard University Press.

Tóth, L., Kovács, G., Van Compernolle, D. (2018). A Perceptually Inspired Data Augmentation Method for Noise Robust CNN Acoustic Models. In: Karpov, A., Jokisch, O., Potapova, R. (eds) *Speech and Computer*. SPECOM 2018

Tottie, G. (1991) Conversational style in British and American English: The case of backchannels. In Aijmer, K. and Altenberg, B. (Eds.) English corpus linguistics. London: Longman. pp.254-271.

Truong, K. P., Poppe, R., and Heylen, D. (2010). A rule-based BC prediction model using pitch and pause information. *In Eleventh Annual Conference of the International Speech Communication Association*.

Trujillo, J., Özyürek, A., Holler, J. et al. Speakers exhibit a multimodal Lombard effect in noise. *Sci Rep 11*, 16721 (2021). https://doi.org/10.1038/s41598-021-95791-0

Türker, B. B., Erzin, E., Yemez, Y., & Sezgin, T. M. (2018, January). Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions. In *Interspeech* (pp. 1741-1745).

Watzlawick, P., Beavin, J. and Jackson, D. (1967) Pragmatics of Human Communication. W. W. Norton: New York.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics*, 32(8), 1177-1207.

Ward, N. G., & Vega, A. (2009, November). Towards the use of inferred cognitive states in language modeling. *In 2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 323-326). IEEE.

White, S. (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, 18(1), 59-76.

Viikki, O., and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3), 133-147.

Ward, N. (1996, October). Using prosodic clues to decide when to produce back-channel utterances. *In Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP'96 (Vol. 3, pp. 1728-1731). IEEE.

Ward, N., 1997. Inferring processing times from a corpus of conversations. Handout for the *First Conference on Computational Psycholinguistics*.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Roux, J. L., Hershey, J. R., & Schuller, B. (2015, August). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. *In International conference on latent variable analysis and signal separation (pp. 91-99)*. Springer, Cham.

White, S. (1989) Backchannels across cultures: A study of Americans andJapanese. *Language in Society* 18: pp.59-76.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. *In Proceedings* *of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Wong, D., & Kruger, H. (2018). Yeah, yeah yeah or yeah no that's right: A multifactorial analysis of the selection of backchannel structures in British English. *In Corpus Approaches to Contemporary British Speech* (pp. 120-156). Routledge.

Xu, Y., & Reitter, D. 2017. Spectral analysis of information density in dialogue predicts collaborative task performance. *In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics*. Vancouver, Canada

Xu, Y., and Reitter, D. (2018). Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170, 147-163.

Yngve, V. H. (1970). On getting a word in edgewise. In Chicago Linguistics Society, 6th Meeting, 1970 (pp. 567-578).

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In A. Papafragou, D. Grodner, D. Mirman, & 690 J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2093–2098). Austin, TX: Cognitive Science Society.

Zhang, Y., Sun, S., Galley, M., Chen, Y. C., Brockett, C., Gao, X., ... and Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536.*

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, 149(6), 3910-3916.

Appendices

Section I: Supplementary materials on feature extraction

Table 1 *Summary of BC inviting cues*

Modality	Category	Features	Studies
Verbal	Word embedding	Word2vec	Ruede et al., 2017
	Part of speech	POS tags	Cathcart et al. 2003
	Lexical information	Polarity & concreteness	Boudin et al., 2021
	Verbose	Verbose	Park et al., 2017; Gravano and Hirtschberg, 2009
Vocal	Pitch	Low pitch region	Murray et al 2021; Ruede et al., 2017; Ward and Tsukahara, 2000
		Pitch slope	Boudin et al., 2021; Morency et al, 2010
		Pitch variation	Gravano and Hirschberg, 2009; Jain et al., 2021; Moubayed, 2009
	Voice quality	Noise-to-harmonic ratio	Gravano and. Hirschberg, 2009
		Jitter; Shimmer	Levitan et al., 2011
	Energy /intensity	Intensity variation	Gravano and. Hirschberg, 2009; Ruede et al., 2017
	Lengthened vowels	F1, F2, F3 frequency	Park et al.,2017; Ward and Tsukahara, 2000
	Ceptral	MFCC	Goswami et al., 2020; Jain et al., 2021; Murray et al 2021; Ruede et al., 2021
	Pause	Energy	Boudin et al., 2021; Cathcar et al., 2003; Jain et al., 2021; Park et al., 2017
Visual	(Mutual) gaze	Blink; eye movement	Jain et al., 2021; Tuong et al., 2011
		pupil dilation	Goswami et al., 2020
	Head movement	Translational/rotational velocity; acceleration	Boudin et al., 2021; Moubayed, 2009; Murray et al 2021; Jain et al., 2021
	Posture	Forward; backforward	Jindal et al., 2021
	Mouth	Smile	Moubayed, 2009

Table 2

Nonverbal behaviors annotations and tags

Category	Tag	Annotated Feature	Explanation
Turn	Speech	IPUs	
Speech Function	Feedback	Vocal feedback and BCs	Short IPUs functioning as feedback/BC
	Response	Short responses to questions	Short IPUs used as responses
Gaze	LS	Looking at the screen	Looking at the screen = looking at the other participant
	LA	Looking away	When it's not LS, it is LA
Head	NodR	Nods	Nods as response
	NodF	Nods	Nods as feedback/BC
	Nod	Nods	Non communicative Nods
	HShake	Headshakes	Both communicative and non-communicative (function is annotated in using FuncH tier when it is used for responding)
Eyebrow	Raised	Eyebrow raising	
	Frown	Frowning	
Mouth	S1	Smile	Smile with closed mouth
	S2	Smile	Smile with open mouth
	Laugh	Laugh	Laugh (with the sound)
Posture	Forward	Leaning forward	With respect to the rest position
	Backward	Leaning backward	With respect to the rest position

*This table is from Bodur (2021)

Table 3

Average gamma scores quantifying inter-rater reliability between two annotators using 20% of the corpus. Ranges indicate lowest and largest gamma in the videos annotated in each age group.

Features	**Children**		**Adults**	
	Categorization	**Segmentation**	**Categorization**	**Segmentation**
Gaze	0.93 [0.85, 0.99]	0.68 [0.63, 0.73]	0.98 [0.94, 1.00]	0.76 [0.61, 0.88]
Mouth_Smile	0.84 [0.66, 1.00]	0.55 [0.32, 0.75]	0.96 [0.94, 1.00]	0.58 [0.42, 0.70]
Mouth_Laugh	0.81 [0.58, 1.00]	0.67 [0.49, 0.86]	0.99 [0.94, 1.00]	0.79 [0.64, 0.87]
Head_Shake	0.99 [0.94, 1.00]	0.69 [0.39, 0.89]	0.94 [0.87, 1.00]	0.71 [0.48, 0.83]
Head_Nod	0.86 [0.65, 1.00]	0.57 [0.47, 0.78]	1.00 [1.10, 1.00]	0.57 [0.46, 0.68]
Posture_Forward	0.81 [0.67, 1.00]	0.50 [0.33, 0.80]	0.90 [0.79, 1.00]	0.63 [0.49, 0.88]
Posture_Backward	0.86 [0.74, 0.94]	0.52 [0.33, 0.68]	0.94 [0.83, 1.00]	0.67 [0.46, 0.91]
Eyebrow_Raised	0.82 [0.77, 0.94]	0.50 [0.43, 0.56]	0.92 [0.88, 0.97]	0.66 [0.57, 0.77]
Eyebrow_Frown	0.79 [0.71, 0.86]	0.52 [0.37, 0.68]	0.66 [0.47, 0.77]	0.49 [0.45, 0.53]

*This table is from Bodur et al., 2021

Table 4

Example transaction annotations of English dialogue

#G	#L	Speaker	Utterance	H(S)	H(S|C)
6	1	Initiator	Mm-hmm. Mm, animal?	15.001	14.256
7	2	Responder	It is not an animal.	14.021	19.150
8	3	Initiator	it's an animal? Okay. Mm, so it's a human?	14.400	18.681
9	4	Responder	No, it's not a human either.	14.756	17.502
10	5	Initiator	It's a human?	13.783	18.060
6	6	Responder	It's not a human.	13.917	18.745
			…		
8	8	Responder	Yeah. Okay. Do you wanna do one more and you, you, you think of the word and I guess?	17.911	14.247
9	9	Initiator	Okay. Uh, and, uh, wait, I'm, I'm thinking for one.	18.773	15.333
1	1	Initiator	Taking one?	19.343	14.073
2	2	Responder	I'm, I'm thinking. I'm	13.430	16.306
3	3	Initiator	Okay. Real, real quick because then I have to, I, I, I wanted to ask you a question. Well, I guess I can ask you now...	22.530	14.385

Table 5 *Summary of BC modeling studies*

Model	reference	Dataset				context	sr	Extracted features		
		Size	age	No.	task			visual	acoustic	verbal
LSTM	Murray et al, 2021	136.08	AA	12	T	2s	20ms	head pose	F0; MFCC	NA
	Jain et al, 2021	702	AA	38	T	3s	NA	Gaze; head	F0; MFCC; energy	NA
	Ruede et al., 2017	15600	AA	NA	T	32 ms	10ms	NA	Pitch; MFCC; energy	NA
	Goswami et al. 2020	75	CC	18	T	3s	30HZ	Gaze; head	F0; MFCC; energy	NA
Random Forest	Moubayed, 2009	NA	AA	NA	T	NA	NA	Smile; head	pitch	NA
Probabilistic	Morency et al, 2010	NA	AA	104	T	NA	30HZ	gaze	Pitch; Intensity loudness	unigram
HMM	Solorio et al.,2006	110	NA	NA	S	NA	NA	NA	Pitch; energy	NA
rule-based	Boudin et al. 2021	420	NA	NA	S	NA	event	Nods; laugh; smiles	tone	POS polarity
	Park et al,2017	NA	CC	18	T	NA	NA	Gaze; eyebrow	pitch, energy, pause	NA
	Ward et al., 2000	80	CC	24	T	NA	NA	NA	pitch range	NA

Note: size(min.); age(AA:adult-adult conversations; CA: child-caregiver conversations)
sr: sampling rate; task: (T: task-oriented; S: simultaneous)

Section II: Supplementary materials on preliminary analysis of information entropy

Q1 Is information transmission rate constant?

We hypothesize that the UID principle will be more visible at the transaction level, where the context is more coherent in content compared with the dialogue level. However, it is still unclear whether this principle holds for both child-caregiver conversations. On the one hand, children are likely to be constrained by language proficiency and the ability to infer others' communicative intent. As a result, their prediction of parents' interpretation differs correspondingly. On the other hand, as parents tend to adjust their expressions to adapt to children's production. The general information transmission rate may stay uniform, but lower than the adult-adult controls.

Following Giulianelli and Fernandez (2021), I fitted linear mixed-effects regression models on child-caregiver and adult-adult conversation data with the decontextualized information content H(S) as the response variable and the utterance position (either within the dialogue or topical units), utterance length, data set (CA vs. AA) coded with sum contrast scaled to values of -0.5 for adult-adult conversations and +0.5 for child-caregiver conversations, the interactions between data set and utterance position, and between data set and utterance length as predictors, with a random intercept grouped by distinct dialogues . Considering the potential confounding effect of utterance length (Keller, 2004; Xu and Reitter, 2018), utterance length was added as another predictor. The whole model was specified as

*entropy ∼ position * conversation type + length * conversation type + (1 | interlocutor)*

We repeated the same procedure to fit models of the contextualized information content H(S|C), and the mutual information I (S; C) as response variables. The results of the linear mixed-effect models are summarized in Table 6 below.

For the contextualized information entropy H(U|C), I didn't find significant effect of sentence position (global: $\beta = 0.014$, $p = 0.618$; local: $\beta = 0.041$, $p = 0.342$) in both child-caregiver and adult-adult conversations, suggesting that the information transmission rate remained constant regardless of interlocutors. For the decontextualized information entropy H(U), I did not find a significant positive effect of the whole dialogue utterance position on information content ($\beta = 0.032$, $p = 0.176$), indicating that there was no increasing trend on the whole dialogue level. In contrast, decontextualized information content increased with utterance position ($\beta = 7.779e{-}03$, $p < 0.05$) in a topically compact context. A further analysis indicated that such increase was significantly higher in adult-adult dialogues than child-caregiver dialogues. For the context informativeness I (U; C), there was no significant increase with the unfolding of the whole dialogue or topical unit.

In summary, the analysis above empirically confirms that the ERC principle holds on both dialogue level regardless of interlocutors. H(U) and I(U; C), however, do not always increase together, and when they do, they grow at a different rate. The regression coefficients are rather small but comparable to those found in prior work (Qian and Jaeger, 2011; Xu and Reitter, 2018; Giulianelli et al., 2021)

Table 6

Summary of regression model fitted to information density

		Fixed effects			Random effects
		Estimate	**SE**	**Pr(>|t|)**	
H(U|C)	Intercept	0.051	0.265	0.848	0.624
	Dataset	-0.053	0.365	0.887	
	Global position	0.014	0.020	0.618	
	Utterance length	-0.085	0.014	< 0.001	
	Dataset * Global position	-0.016	0.025	0.520	
	Dataset * Utterance length	0.030	0.023	0.195	
	Intercept	0.050	0.265	0.853	0.624
	Dataset	-0.051	0.365	0.890	
	Local position	0.041	0.016	0.342	
	Utterance length	-0.083	0.014	< 0.001	
	Dataset * Local position	0.027	0.023	0.232	
	Dataset * Utterance length	-0.061	0.023	< 0.01	
H(U)	Intercept	0.031	0.082	0.714	0.963
	Dataset	-0.030	0.112	0.793	
	Global position	0.032	0.031	0.176	
	Utterance length	0.077	0.022	< 0.001	
	Dataset * Global position	-0.013	0.038	0.724	
	Dataset * Utterance length	0.158	0.035	< 0.001	
	Intercept	2.154e-02	8.260e-02	0.797	0.961
	Dataset	-1.821e-02	1.133e-01	0.874	
	Local position	**7.779e-03**	**2.516e-02**	**< 0.05**	
	Utterance length	8.095e-02	2.169e-02	< 0.001	
	Dataset * Local position	**6.990e-02**	**3.482e-02**	**< 0.05**	
	Dataset * Utterance length	1.567e-01	3.502e-02	< 0.001	
I (U; C)	Intercept	-1.723e-02	1.971e-01	0.931	0.802
	Dataset	1.844e-02	2.715e-01	0.947	
	Global position	1.522e-02	2.576e-02	0.339	
	Utterance length	1.228e-01	1.840e-02	< 0.001	
	Dataset * Global position	-2.101e-04	3.164e-02	0.995	

Dataset * Utterance length	8.820e-02	2.947e-02	< 0.001	
Intercept	-0.023	0.195	0.905	0.801
Dataset	0.027	0.269	0.922	
Local position	-0.024	0.021	0.089	
Utterance length	0.124	0.018	< 0.001	
Dataset * Local position	0.097	0.029	< 0.001	
Dataset * Utterance length	0.088	0.029	< 0.01	

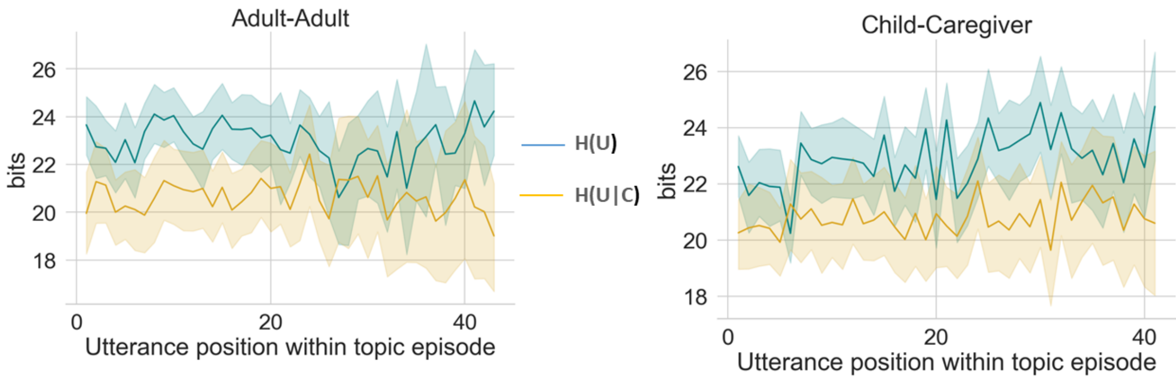


Figure 1 Decontextualised information content H(U), contextualised information content H(U|C), and context informativeness I(U;C) against utterance position within the topic episode. Bootstrapped 95% confidence intervals

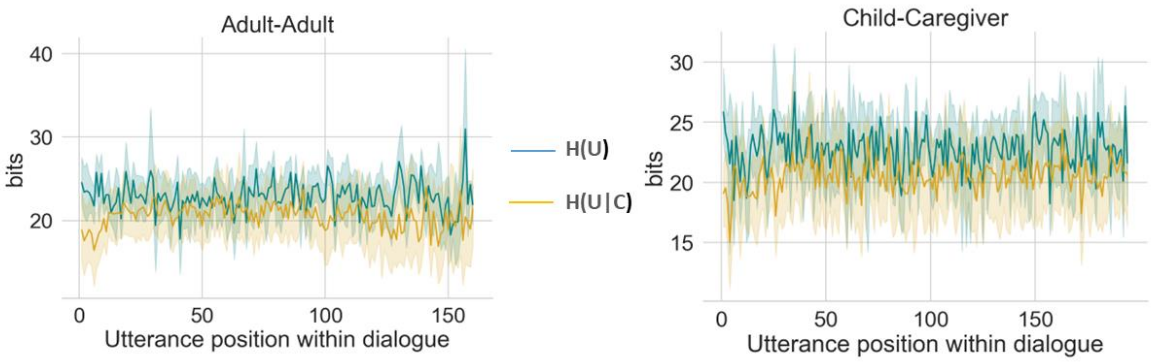


Figure 2 Decontextualised information content H(U), contextualised information content H(U|C), and context informativeness I(U;C) against utterance position within the whole dialogue. Bootstrapped 95% confidence intervals

Q2 Does utterance information converge between speaker roles?

Following Xu and Reitter(2018)'s study, I tested whether there is a converging trend of different speakers roles for both child-caregiver and adult-adult conversations. Speakers' roles in each topic episode were manually annotated as elaborated in Section 3.2.

Two linear mixed effects models were fitted for child-caregiver and adult-adult conversations respectively with the decontextualised information content H(U) as the response variable and the utterance position within topical units as well as the interactions between speaker role and utterance position, as predictors with a random intercept grouped by distinct dialogues.

*entropy ~ position * conversation type + length * conversation type + (1 | interlocutor)*

As Figure 3 shows, there was a significant interaction effect between speaker roles and utterance position for both types of conversations in child-caregiver conversations ($\beta = 0.115$, $p < 0.01$) which was in aligned with the results in Xu and Reitter(2018) but not in adult-adult conversations ($\beta = -0.087$, $p = 0.087$).

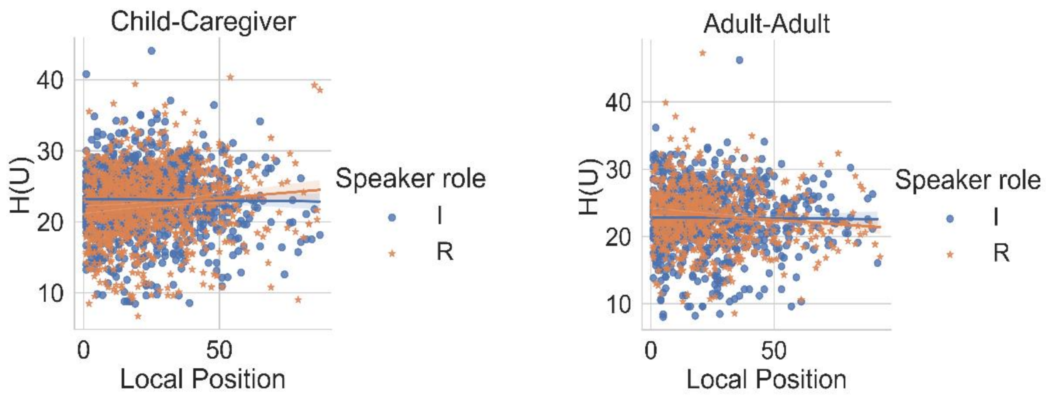


Figure 3 Utterance information against the relative utterance position within topic episodes, grouped by speaker roles (topic initiator vs. responder). Bootstrapped 95% confidence bands.

Section III: Supplementary materials on ranked features in Chapter 4

Table 7

Ranked features on different (combinations of) modalities after RFE

7a *Ranked features of verbal modality after RFE*

No.	Child	Caregiver	Adult1	Adult2
1	'NOUN'	'NOUN'	'ADP'	'INTJ'
2	'DET'	'VERB'	'NOUN'	'VERB'
3	'VERB'	'ADJ'	'ADV'	'ADV'
4	'PROPN'	'DET’	'DET'	'NOUN'
5	'CCONJ'	'CCONJ'	'INTJ'	'PRON'
6	'INTJ'	‘AUX'	'VERB'	'DET'
7	'ADJ'	'Surprisal'	'NUM'	'PROPN'
8	'ADP’	'PROPN'	'Surprisal'	'ADP'
9	'AUX'	'PRON'	'CCONJ'	'ADJ'
10	'SCONJ'	'ADV'	'SCONJ'	'CCONJ'
11	'Surprisal'	'ADP'	'ADJ'	'Surprisal'
12	'ADV'	'INTJ'	'PROPN'	'SCONJ'
13	'NUM'	'SCONJ'	'AUX'	'NUM'
14	'PRON'	'NUM'	'PRON'	'AUX'
15	'PART'	'PART'	'PART'	'PART'

7b

Ranked features of visual modality after RFE

No.	Child	Caregiver	Adult1	Adult2
1	'Gaze_no'	'Smile_no'	'Gaze_no'	'Hshake_dur'
2	'Laugh'	'Raised_no'	'Frown_no'	'Nod_dur'
3	'Laugh_dur'	'Hshake_dur'	'Backward_dur'	'HShake_no'
4	'Smile_dur'	'HShake_no'	'Smile_dur'	'Raised_no'
5	'Raised_no'	'Nod_no'	'HShake_no'	'Smile_no'
6	'Frown_dur'	'Frown_no'	'Backward_no'	'Forward_dur'
7	'Smile_no'	'Frown_dur'	'Laugh_dur'	'Nod_no'
8	'Raised_dur'	'Raised_dur'	'Gaze_dur'	'Raised_dur'
9	'Forward_no'	'Backward_no'	'Forward_no'	'Backward_dur'
10	'HShake_no'	'Nod_dur'	'Forward_dur'	'Forward_no'
11	'Hshake_dur'	'Smile_dur'	'Laugh'	'Gaze_no'
12	'Backward_no'	'Laugh_dur'	'Raised_no'	'Frown_dur'
13	'Nod_dur'	'Gaze_dur'	'Nod_no'	'Backward_no'
14	'Nod_no'	'Backward_dur'	'Nod_dur'	'Laugh'
15	'Forward_dur'	'Forward_dur'	'Smile_no'	'Frown_no'
16	'Frown_no'	'Gaze_no'	'Raised_dur'	'Gaze_dur'
17	'Backward_dur'	'Laugh'	'Hshake_dur'	'Smile_dur'
18	'Gaze_dur'	'Forward_no'	'Frown_dur'	'Laugh_dur'

7c *Ranked features of vocal modality after RFE*

No.	Child	Caregiver	Adult1	Adult2
1	'loudness_RisingSlope'	'F0semitone_percentile20.0'	'F0semitone_RisingSlope'	'F0semitone'
2	‘VoicedSegmentsPerSec'	'F0semitone_amean'	'MeanVoicedSegmentLengthSec	'F0semitone_RisingSlope'
3	‘Mean Voiced Segment Length Sec'	‘Mean Unvoiced Segment Length'	'Voiced Segments Per Sec	'loudness_std'
4	'F0semitone_amean'	‘HNR’	'F0semitone_amean'	'F0semitone_percentile20.0'
5	'loudness_ amean'	'loudness_ meanRisingSlope'	‘MeanUnvoicedSegmentLength'	‘VoicedSegmentsPerSec'
6	'loudness_Falling Slope'	‘MeanVoicedSegmentLengthSec'	'F0semitone_percentile20.0'	'F0semitone_amean'
7	'F0semitone_FallingSlope'	'jitter’	‘shimmer’	‘shimmer’
8	'F0semitone_RisingSlope'	‘VoicedSegmentsPerSec'	'loudness_ amean'	'LocaldB_amean'
9	'loudness_ std'	‘shimmer’	'F0semitone_meanFallingSlope'	'loudness_FallingSlope'
10	'F0semitone_std'	'F0semitone_meanFallingSlope'	'loudness_std'	'F0semitone_FallingSlope'
11	‘Mean Unvoiced Segment Length'	'F0semitone_std'	'loudness_FallingSlope'	'jitter’
12	‘shimmer’	'loudness_Falling Slope'	'F0semitone_std'	'MeanUnvoicedSeg Length'
13	'F0semitone percentile20'	'loudness_RisingSlope'	'loudness_ meanRisingSlope'	'loudness_ meanRisingSlope'
14	‘HNR’	'loudness_std'	‘HNR’	'dBACF_sma3nz_amean'
15	'jitter’	'loudness_ amean'	'jitter’	'MeanVoicedSegLengthSec'

7d *Ranked features of vocal-verbal modality combination after RFE*

No.	Child	Caregiver	Adult1	Adult2
1	'loudness_ amean'	'F0semitone_percentile20'	'loudness_ amean'	'F0semitone_amean''
2	'loudness_RisingSlope'	'F0semitone_amean''	'ADP'	'DET'
3	'DET'	'MeanUnvoicedSegmentLength'	'INTJ'	'VoicedSegmentsPerSec’
4	'F0semitone_amean'	'MeanVoicedSegmentLengthSec'	'loudness_RisingSlope'	'F0semitone_std'
5	'F0semitone_percentile20'	'loudness_RisingSlope'	'PRON'	'loudness_RisingSlope'
6	'F0semitone_FallingSlope'	'VERB'	'VoicedSegmentsPerSec'	'loudness_mean'
7	'SCONJ'	'loudness_ amean'	'F0semitone_std'	'MeanUnvoicedSegmentLength'
8	'Surprisal'	'ADJ'	‘shimmer’	‘HNR’
9	'loudness_meanFallingSlope'	'NOUN'	'NOUN'	'PRON'
10	'MeanVoicedSegmentLengthSec'	'F0semitone_FallingSlope'	'PROPN'	'MeanVoicedSegmentLengthSec'
11	'loudness_sma3_std’	'CCONJ'	'CCONJ'	‘shimmer’
12	'VERB'	'PRON'	'AUX'	'VERB'
13	'VoicedSegmentsPerSec'	'loudness_std'	'NUM'	'Surprisal'
14	'PROPN'	'F0semitone_amean''	'SCONJ'	'F0semitone_FallingSlope'
15	'NOUN'	'F0semitone_std'	'jitter’	'ADP'
16	'F0semitone_std'	'F0semitone_RisingSlope'	'VERB'	'PROPN'
17	'shimmer’	'ADP'	'F0semitone_FallingSlope'	'INTJ'
18	'CCONJ'	'Surprisal'	'Surprisal'	'NOUN'
19	'PRON'	'DET'	'loudness_std'	'AUX'

20	'ADV'	'AUX'	‘HNR’	'F0semitone_percentile20.0'
21	'AUX'	'ADV'	'ADJ'	'loudness_FallingSlope'
22	'ADP'	'jitter’	'MeanUnvoicedSeg len’	'F0semitone_RisingSlope'
23	'ADJ'	'PROPN'	'F0semitone_RisingSlope'	'jitter’
24	'F0semitone_RisingSlope'	'shimmer’	'ADV'	'ADJ'
25	'MeanUnvoicedSegmentLength'	'SCONJ'	'F0semitone_amean''	'CCONJ'
26	'HNR’	'loudness_meanFallingSlope'	'DET'	'ADV'
27	'INTJ'	‘HNR’	'loudness_FallingSlope'	'loudness_std'
28	'jitter’	'NUM'	'MeanVoicedSegLenSec'	‘SCONJ'
29	'NUM'	'INTJ'	'F0semitone_percentile20'	'NUM
30	'PART'	'PART'	'PART'	'PART'

7e *Ranked features of all modality combination after RFE*

No.	Child	Caregiver	Adult1	Adult2
1	'loudness_amean'	'loudness_amean'	'MeanUnvoicedSegLength'	'SCONJ'
2	'Nod_dur'	'F0semitone_amean'	'INTJ'	'Raised_dur'
3	'Nod_no'	'F0semitone_percentile20'	'PRON'	'F0semitone_RisingSlope'
4	'loudness_RisingSlope'	'MeanVoicedSegmentLengthSec'	'F0semitone_amean'	'Nod_dur'
5	'MeanVoicedSegmentLengthSec'	'F0semitone_RisingSlope'	'Hshake_dur'	'VoicedSegmentsPerSec'
6	'Hshake_dur'	'VERB'	'VoicedSegmentsPerSec'	'NOUN'
7	'INTJ'	'F0semitone_std’	'Laugh_dur'	'Smile_dur'
8	'VoicedSegmentsPerSec'	'Nod_dur'	'PROPN'	'MeanUnvoicedSegmentLength'
9	'Smile_no'	'ADJ'	'F0semitone_percentile20'	'PROPN'
10	'Raised_dur'	'SCONJ'	'shimmer’	'shimmer’
11	'Backward_dur'	'CCONJ'	'CCONJ'	'ADP'
12	'MeanUnvoicedSegmentLength	'F0semitone_RisingSlope'	'loudness_amean'	'Gaze_dur'
13	'Frown_dur'	‘Laugh'	'Forward_dur'	'AUX'
14	'Gaze_dur'	'Gaze_dur'	'Raised_no'	'loudness_FallingSlope'
15	'HNR’	'Forward_dur'	'AUX'	'Frown_no'
16	'Surprisal'	'Forward_no'	'ADV'	'Forward_no'
17	'Laugh_dur'	'Backward_no'	'HNR’	'Backward_dur'
18	'PROPN'	'Hshake_dur'	'loudness_FallingSlope',	'Laugh'
19	'F0semitone_RisingSlope'	'PRON'	'Raised_dur'	'Hshake_dur'

20	'F0semitone_FallingSlope'	'NOUN'	'NOUN'	'loudness_amean'
21	'F0semitone_percentile20'	'Frown_dur'	'Nod_dur'	'Frown_dur'
22	'DET'	'loudness_FallingSlope'	'loudness_RisingSlope'	'jitter’
23	'ADJ'	'Smile_no'	'ADP'	'Gaze_no'
24	'SCONJ'	'F0semitone_FallingSlope'	'DET’	'Surprisal'
25	'shimmer'	'Raised_dur'	'Laugh'	'Raised_no'
26	'CCONJ'	'Backward_dur'	'Backward_dur'	'INTJ'
27	'AUX'	'shimmer'	'NUM'	'F0semitone_FallingSlope'
28	'ADP'	'Laugh_dur'	'Smile_dur'	‘HNR’
29	'Gaze_no'	'ADP’	'Gaze_no'	'NUM'
30	'ADV'	'loudness_std'	'loudness_std'	'MeanVoicedSegmentLengthSec'
31	'VERB',	‘AUX	'Gaze_dur'	'DET'
32	'jitter’	VoicedSegmentsPerSec'	'Nod_no'	'ADV'
33	'loudness_FallingSlope',	'Surprisal'	'F0semitoneRisingSlope'	'F0semitone_std'
34	'PRON'	'Smile_dur'	jitter	'Backward_no'
35	'HShake_no'	'Raised_no'	'F0semitone_FallingSlope'	'Smile_no'
36	'Raised_no'	‘ADV'	'SCONJ'	'F0semitone _amean'
37	'Laugh'	'HNR’	'Frown_no'	'VERB'
38	'Forward_dur'	'PROPN'	'Frown_dur'	'ADJ'
39	'NOUN'	'Gaze_no'	'F0semitone_std’	'Nod_no'
40	'F0semitone_amean'	'Nod_no'	'Smile_no'	'Forward_dur'

41	'Smile_dur'	'HShake_no'	'MeanVoicedSegLenSec'	'PRON'
42	'Forward_no'	'DET'	'Forward_no'	'F0semitone _percentile20'
43	'Backward_no'	‘MeanUnvoicedSegmentLength'	‘VERB'	'HShake_no'
44	'loudness_std'	jitter	'Backward_no'	'CCONJ'
45	'F0semitone_std'	'Frown_no'	'HShake_no'	'Laugh_dur'
46	'NUM'	'INTJ'	'Surprisal'	'loudness_std'
47	'Frown_no'	'NUM'	'ADJ'	'loudness_RisingSlope'
48	'PART'	'PART'	'PART'	'PART'

7d *Ranked features of visual-verbal modality combination after RFE*

No.	Child	Caregiver	Adult1	Adult2
1	'NOUN'	'Smile_no'	'ADP'	'NOUN'
2	'Nod_no'	'VERB'	'PRON'	'INTJ'
3	'Nod_dur'	'ADJ'	'Laugh_dur'	'SCONJ'
4	'Frown_dur'	'Nod_no'	'CCONJ'	'Laugh_dur'
5	'Gaze_no'	'Raised_dur'	'VERB'	'PRON'
6	'Raised_no'	'Backward_dur'	'Hshake_dur'	'Nod_no'
7	'SCONJ'	'Backward_no'	'Raised_no'	'Backward_dur'
8	'ADP'	'Smile_dur'	'Backward_dur'	'Raised_dur'
9	'Laugh_dur'	‘AUX'	'SCONJ'	'AUX'
10	'DET'	‘ADP'	'Gaze_dur'	'ADP'
11	'HShake_no'	'Gaze_no'	'Forward_dur'	'Forward_no'
12	'AUX'	'Laugh'	'Nod_no'	'CCONJ'
13	'Laugh'	'DET'	'Smile_dur'	'PROPN'
14	'Backward_dur'	'HShake_no'	'PROPN'	'VERB'
15	'INTJ'	'CCONJ'	'ADJ'	'Frown_no'
16	'Forward_no'	'PRON'	'Raised_dur'	'Frown_dur'
17	'Smile_dur'	'NOUN'	'Gaze_no'	'Smile_dur'
18	'CCONJ'	'Forward_dur'	'INTJ’	'NUM’
19	'Hshake_dur'	'ADV'	'Surprisal'	'Raised_no'
20	'Gaze_dur'	'PROPN'	'NOUN’	'Hshake_dur'
21	'PROPN'	'Surprisal'	'Forward_no'	'HShake_no'
22	'ADV'	'Nod_dur'	'Nod_dur'	'Laugh'
23	'Smile_no'	'Gaze_dur'	'NUM'	'ADJ'
24	'PRON'	'Frown_no'	'HShake_no'	'Forward_dur'
25	'ADJ'	'Frown_dur'	'Frown_no'	'Backward_no'
26	'NUM'	‘SCONJ'	'Laugh'	'Nod_dur'
27	'Surprisal’	'Raised_no'	'Frown_dur'	'Gaze_no'
28	'Forward_dur'	'Laugh_dur'	'Backward_no'	'DET'
29	'Frown_no'	'Hshake_dur'	'DET'	'Surprisal'
30	'Backward_no'	'Forward_no'	'ADV'	'Smile_no'
31	'Raised_dur'	'NUM'	'Smile_no'	'Gaze_dur'
32	'VERB'	'INTJ'	'PART'	'ADV'
33	'PART'	'PART'	'AUX'	'PART'