Predicting Backchannel Signaling in Multimodal Child-Caregiver Conversations

Anonymous WoCBU submission

ABSTRACT

Conversation requires cooperative social interaction between interlocutors. In particular, active listening through backchannel signaling (hereafter BC) i.e., showing attention through verbal (short responses like "Yeah") and non-verbal behaviors (e.g. smiling or nodding) is crucial to manage the flow of a conversation and requires sophisticated coordination skills. How does BC develop in childhood? Previous studies were either conducted in highly controlled experimental settings or relied on qualitative corpus analysis, which do not allow for a proper understanding of children's BC development, especially in terms of its collaborative/coordinated use. This paper aims at filling this gap using a machine learning model that learns to predict children's BC production based on the interlocutor's inviting cues in child-caregiver naturalistic conversations. By comparing BC predictability across children and adults, we found that, contrary to what have been suggested in previous in-lab studies, children between the ages of 6 and 12 can actually produce and respond to backchannel inviting cues as consistently as adults do, suggesting an adult-like form of coordination.

KEYWORDS

cognitive development; language acquisition; conversation; backchannel; nonverbal

1 INTRODUCTION

Backchannel [41] is an important conversational mechanism that interlocutors use — when in the listening mode — to provide non-intrusive signals to the speaker such as "uh-huh" and/or nonverbal gestures such as head nods. BC is crucial for a coordinated conversation as it is often used to signal attention and understanding (or lack thereof) while allowing the speaker to retain the main channel and make the necessary adjustments toward achieving mutual understanding [7].

This paper deals with the development of BC in middle childhood. Inspection of the literature shows that most previous studies on BC perception or/and production have focused on the preschool period, i.e., the period in childhood generally up to around 5 years old [e.g., 30, 31, 34]. However, much of children's conversational skills continue developing through middle childhood (between 6 and 12 years old) as their social interactions become more sophisticated and their socio-cognitive skills undergo developmental changes [9].

However, little is known about children's BC signalling at this crucial stage of development. The few existing studies [e.g., 11, 23] had indicated that children's BC skills were still limited and far from having achieved adult-like maturity. However, a more recent work by [3] argued that these previous studies might have under-estimated children's competence due probably to the — less than optimal — context of data collection (e.g., having an adult stranger as the child's conversational partner, using a scripted dialog, holding the conversation in the lab as opposed to children's homes) since research on adults has shown that the way people tend to use various conversational mechanisms, including backchannels, depends on the context of the conversation [10].

They collected new data in a context that was supposed to be more natural to children, i.e., a context similar to how they communicate spontaneously in daily life. As a data acquisition method, they capitalized on the recent increase in familiarity and use of online video calls by children — due to Covid-19 pandemic — to record video calls between children and their caregivers at home while playing a simple, intuitive verbal game. In contrast to previous studies, [3] found that children aged 6 to 11 years old produced BC at a rate that was comparable to that of adults.

However, in order for children to achieve adult-level mastery, they should be able, not only to produce BC at a reasonable frequency, but also learn *when* to produce them. In other words, even if children's rate of production is similar to that of adults, they may not necessarily be as good in terms of timing these productions with the speaker's inviting cues. This *collaborative* aspect of BC production is crucial for conversation to go smoothly. Indeed, BCs are not as effective if produced randomly; they may even be perceived as distracting or interrupting [e.g., 30].

The current study

The main goal of this work is to compare children's collaborative BC skills to those displayed by adults. This comparison allows us to better characterize the developmental trajectory: Are children in middle childhood already on par with adults? Or are their BC coordination skills rather immature and still undergoing developmental change?

We investigate children's level of BC mastery in terms of collaborative production and elicitation of BC, using the same corpus of conversations analysed in [3]. When children are in the listening mode, we measure their ability to react consistently to the speaker's inviting cues. These cues can be multimodal and may involve changes in intonation, gaze, gesture, and/or sentence structure. In turn, when in the speaking role, we measure children's

ability to offer inviting cues for BC to their interlocutor using reliable multimodal signaling.

To this end, we use research methods that are not typical in experimental developmental psychology. Typical controlled in-lab designs [e.g., 23] have used scripted dialogues where the cues to BC are predesignated *top-down* by the experimenter. Here, we focus on relatively naturalistic/spontaneous child-caregiver dialog data which requires using rather *bottom-up* tools in order to quantify the degree to which the interlocutors capitalize on each other's cues to produce BC collaboratively.

We borrow techniques from the literature on dialog systems. While our goal here is not to engineer an artificial dialog agent, this literature is rich with methods and insights that are useful to address research questions on child conversational development. In particular, we instantiate collaborative BC dynamics in child-caregiver conversations in a predictive framework: We train a model that learns how to predict the listener's BC based on the speaker's inviting cues (e.g., making a small pause mid-sentence while fixating the interlocutor).

The main intuition is the following: If a trained model can predict the listener's BC based only on the speaker's immediately preceding multimodal communication behavior, this prediction suggests the listener provides BC selectively based on whether or not the speaker's speech contained BC inviting signals, meaning that both interlocutors have been collaborating for the listener to produce BC appropriately.

The earlier models of BC prediction in the field of dialog systems were generally rule-based [2, 18, 30, 36, 38]. Such models have recently been super-seeded in terms of performance by neural networks models [21, 27, 32, 37].

An important feature of a good BC model is its ability to process sequential information. This feature is crucial for learning and testing many important BC inviting cues that are sequential in nature such as the utterance structure and some vocal features (e.g., rising vs. falling intonation) [26]. One neural network model that can handle sequential data, the Long short-term memory (LSTM), have been used in many of state of the art BC modeling [e.g., 21]. It is, therefore, the main model that we use in the current study.

2 METHOD

2.1 Dataset

In this work, we use the Child Conversation corpus [CHICO, 4], consisting of more than 5 hours of Zoom recorded conversations in two conditions: 10 child-caregiver conversations and 10 adult-adult conversations. The second condition is considered here as the "end-state" that children are supposed to reach when their development has matured. Each caregiver participated twice, once with the child and once with another adult, they will be referred to as Adult 1 in the current work (see Results section).

Children were aged between 6 and 11 years old (mean age: 8.7; SD = 1.37). All 10 children were native French speakers. In each conversation, the participants played a word-guessing game for around 10 minutes, changing roles whenever a word was correctly guessed (about 3 to 4 rounds). The game was followed by a more spontaneous conversation. We refer the reader to the original paper [4] for more details about the dataset.

2.2 Models

The broad objective of the LSTM model is to predict the *listener's* BC occurrence based only on the *speaker* multimodal features provided in a time window immediately before the occurrence of the BC. Following previous work [e.g., 21], we used a fixed time window of 2 seconds, which was split into 40 time frames of 50 ms.

In a nutshell, and as illustrated in Figure 1, the network receives, as input, the 40 frames of features extracted from the speaker's data, learn their underlying sequential structure, and then produces a binary output, i.e., BC or no-BC (in a many-to-one setting). Since there are many more frames with no BC than frames with BC, the data would be severely imbalanced. To avoid this issue, we sampled a random set of no-BC frames that was equivalent in size to the BC set (and balanced for each speaker) (see Table 1). The model was trained and tested on this balanced data, leading to intuitively interpretable results.

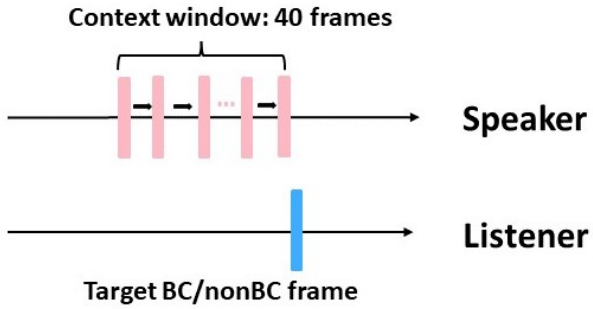

Figure 1: A schematic illustration of the many-to-one mapping in the LSTM model.

We trained and tested several LSTM models to compare the performance of different interlocutors, e.g., a model predicting children's BS vs. a model predicting the caregivers' BC. We also trained models to compare the roles of different modalities, e.g., a model trained using the visual features vs. a model trained using the vocal features.

All models were evaluated using *Leave-one-person-out cross validation* (hereafter LOOCV). For example, a model that predicts children's BC based on the caregivers' cues is trained on all children's BC data except one child. It is then tested on the BC occurrences of the child who was left out in training. This process is repeated with all training/testing configuration.

Bayesian optimization for three hyperparameters (learning rate, dropout, and L2 regularization) were performed for each network configuration using Optuna [1] 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. 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.

SVM Baseline. We compare the results of the LSTM model to a simple Support Vector Machine (hereafter SVM) which we use

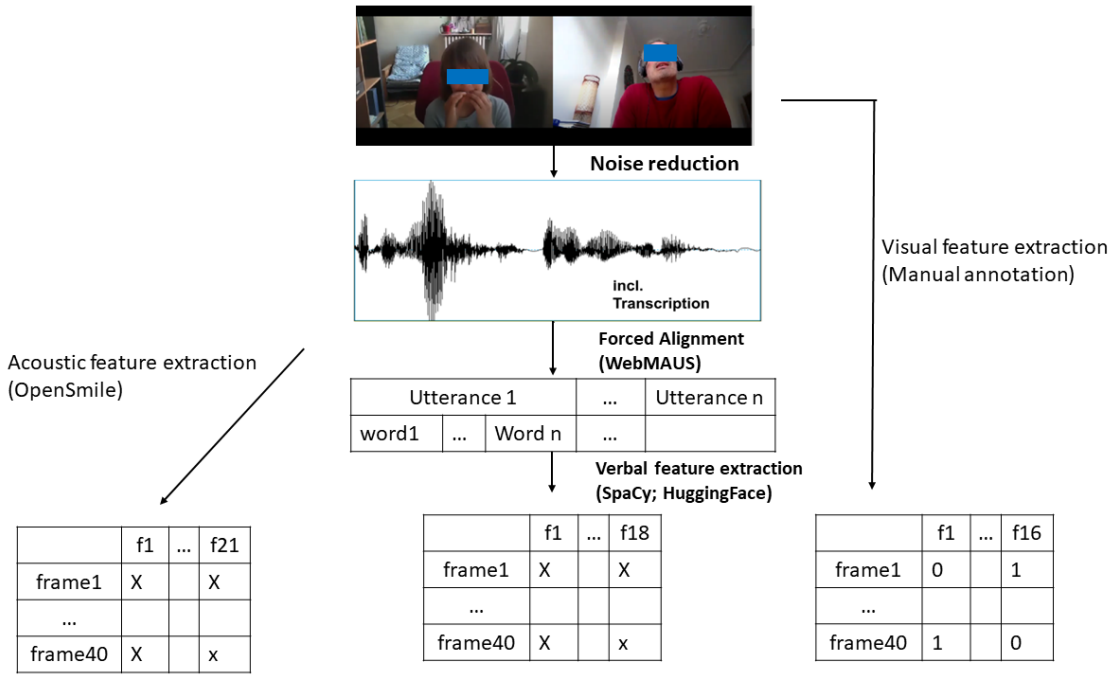

Figure 2: Pipeline of data pre-processing and feature extraction

Table 1: The number of BC in the dataset broken down by the listener category. For balanced training data, we sampled an equivalent number of nonBC, i.e., time frames not containing BC.

Interlocutor	BC	NonBC	Total
All	2841	2841	5682
Child	573	573	1146
Caregiver	454	454	908
Adult1	810	810	1620
Adult2	1004	1004	2008

as a baseline. We selected SVM due to its strength in processing highly dimensional features with no presupposition about feature independence [12]. The input data to the SVM model were vectors with averaged values across the 2-second context window.

2.3 Feature extraction

Figure 2 provides an overview of the pipeline for data pre-processing and feature extraction.

Data pre-processing First, we pre-processed the audio data via non-stationary noise reduction [33]. Second, we used a forced alignment algorithm WebMAUS [24] to automatically derive timestamps at the word level (this step is important for the derivation of the predictive cues belonging to the verbal modality).

Visual features. We included features that have been used in previous work on BC such as head movement (nodding and shaking) [2, 5, 21, 27]), gaze [17, 21, 36], eyebrow movements (raising and frowning [21, 22]), and facial expressions (smiling and laughing [2, 21]). These features were annotated manually and the annotations were available with the dataset (for details, see [4]). We used one-hot encoding where each feature was switched on or off depending on whether or not the feature was present in the current time frame.

Vocal features. We used the Opensmile [14] toolkit to extract the acoustic features from each time frame. We selected a subset of the eGeMAPS that were used in several previous studies, including pitch (variation) [5, 21, 22, 26, 27, 38], Mel-Frequency Cepstral Coefficients (MFCC) [17, 21, 27, 32], voice quality [17, 21, 27, 32], energy [18, 29, 32] and pausal information[5, 6, 21, 30]. To minimize identify-confounding [28] the features were centered and scaled for each participant.

Verbal features. First, we use Part-Of-Speech (POS) as they indicate whether the utterance has ended. This feature is often associated with transitions between discourse units and may elicit communicative responses [5, 6, 36]. We extracted POS tags using SpaCy toolkit [20], resulting in 17 tags for our French dataset. Similar to the visual features, we used one-hot encoding where each POS tag was switched on or off at the frame level. In addition to POS tags, we also used the words' predictability in terms of surprisal, an information-theoretic measure of how unexpected a word is given the prior verbal context. Surprisal has been shown to correlate with communicative efficiency in conversations in previous studies [19] and here we test its ability to predict BC. The word probabilities were extracted using DialoGPT [42], an autoregressive Transformer language model trained and fine tuned on French dialogue materials extracted from films, interviews and theater plays [13, 15, 16, 40]. We relied on HuggingFace's implementation of DialoGPT with default tokenizers and parameters [39]. The word

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

$$Surprisal = -log_2 P(word|context)$$

3 RESULTS

We tested the listeners' collaborative BC production by modeling how an LSTM model would predict their BC as a function of the speakers' inviting cues. In addition, we ran ablation experiments to test the predictive power of each modality. Table 2 shows the mean accuracy scores, with ranges representing the lowest and highest accuracy obtained with different training/testing configurations according to the LOOCV cross-validation.

We can see that the LSTM model performed substantially higher than the SVM baseline (mean accuracies ranging between 0.7 and 0.8), demonstrating and confirming the adequacy of LSTM for BC modeling [21, 32]. When comparing models of children to models of adults, i.e. the main research question of the current work, we can observe that children's BC occurrences are, overall, equally *predictable* as adults', suggesting that children are as responsive as adults to the speaker's inviting cues, achieving, at least quantitatively, an adult-like degree of coordinated behavior.

Further, the fact that caregivers' BCs are also as highly predictable shows, not only that that these caregivers have a consistent BC behavior, but also that their interlocutors (i.e., children) provide consistent inviting cues. This fact suggests that children are also as good as adults in providing reliable inviting cues when in speaking mode.

Finally, when broken down by modality, we found that the three modalities play an equally important role in predicting children's BC as they do in predicting adults' BC.

4 DISCUSSION

This paper investigated collaborative backchanneling behavior in middle childhood through computational modeling of child-caregiver naturalistic interactions. Complementary to previous work applying frequency-based comparison on children-adults' BC usage, here we were interested in the extent to which interlocutors use BC in a collaborative fashion. To this end, we compared children's and adults' BC predictability/responsiveness to the speaker's inviting cues using a neural network (LSTM) that has proven effective in previous research on BC modeling.

Our comparison of models trained on children's and adults' data showed that children's ability to both provide and respond to BC inviting cues was no less consistent than in adults, suggesting that children in middle childhood are not only producing BC at a similar rate as adults do [3] but are also using them in a way that is as predictable and responsive to the interlocutor's cues as fully mature language used do.

Note, however, that while predictability is a sign of coordination, one can imagines scenarios where this is not necessarily the case. For example, children could be consistent in responding to the *wrong* signals/cues, thus being predictable but not collaborative. More thorough examination of the LSTM model is needed to better understand the patterns of children's coordination. Methods dedicated to neural network interpretation would be a good first step.

Table 2: Accuracy scores of our BC predicting models. For each model, we show both the range representing the lowest and highest accuracy obtained with different training/testing configurations according to the LOOCV cross-validation, and the mean accuracy score over this range.

Modality	Listener	SVM	LSTM
visual	Child	0.532[0.485, 0.640]	0.757[0.556, 0.956]
	Caregiver	0.571[0.500, 0.625]	0.759[0.597, 0.971]
	Adult1	0.552[0.500, 0.650]	0.771[0.664, 0.929]
	Adult2	0.536[0.475, 0.620]	0.738[0.529, 0.899]
vocal	Child	0.656[0.529, 0.738]	0.783[0.615, 0.966]
	Caregiver	0.658[0.549, 0.796]	0.734[0.597, 0.971]
	Adult1	0.679[0.593, 0.767]	0.752[0.664, 0.929]
	Adult2	0.689[0.473, 0.767]	0.715[0.493, 0.914]
verbal	Child	0.563[0.500, 0.704]	0.744[0.583, 0.860]
	Caregiver	0.557[0.500, 0.625]	0.806[0.597, 0.971]
	Adult1	0.611[0.500, 0.652]	0.770[0.664, 0.929]
	Adult2	0.626[0.500, 0.708]	0.707[0.493, 0.914]
all	Child	0.652[0.531, 0.733]	0.788[0.613, 0.966]
	Caregiver	0.666[0.537, 0.778]	0.782[0.597, 0.971]
	Adult1	0.672[0.591, 0.725]	0.765[0.654, 0.929]
	Adult2	0.673[0.544, 0.778]	0.720[0.493, 0.908]

Here we could not apply off-the-shelf interpretability algorithms such as SHAP (Lundberg et al., 2017) due to their presupposition of feature independence (a condition that is not met in our data).

More importantly, the current work, like any bottom-up modeling study, remains fundamentally correlational. Indeed, children may well be as predictable as adults. However, this fact does not mean these two populations are using similar mechanism or capitalizing on similar cues. For example, our finding that different modalities predict similarly both children's and adults' BC could be due to the fact that caregivers systematically combine different modalities when inviting BC, and not necessarily to children capitalizing one all modalities. Similarly, our finding that children, when in speaker mode, provide consistent inviting cues (leading to high predictability of caregiver's BC data) is not necessarily due to children explicitly inviting BC from the listening caregiver, but perhaps to the possibility that caregivers produce BC systematically when they observe some specific behavior (e.g., signs of hesitation in children's speech).

Finally, this work focused on children from one culture and with a similar socioeconomic status, which is not enough to draw general conclusion about development. Indeed, many studies indicate that BC occurrence is culturally and contextually specific [8, 25, 35]. For instance, [22] found that the highest educational level of caregivers, gender of both interlocutors and household income influence BC occurrence to a large extent. Future work and data collection should take into account these factors.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] Sames Al Moubayed, Malek Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, J-C Martin, Stanislav Ondas, Catherine Pelachaud, Jérôme Urbain, and Mehmet Yilmaz. 2009. Generating robot/agent backchannels during a storytelling experiment. In *2009 IEEE International Conference on Robotics and Automation*. IEEE, 3749–3754.

[3] Kübra Bodur, Mitja Nikolaus, Abdellah Fourtassi, and Laurent Prévot. 2022. Backchannel Behavior in Child-caregiver Zoom-mediated Conversations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[4] Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. ChiCo: A Multimodal Corpus for the Study of Child Conversation. *ICMI 2021 Companion - Companion Publication of the 2021 International Conference on Multimodal Interaction*, 158–163. https://doi.org/10.1145/3461615.3485399

[5] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, Philippe Blache, and Auriane Boudin Roxane Bertrand Stéphane Rauzy Magalie Ochs Philippe Blache. 2021. A Multimodal Model for Predicting Conversational Feedbacks A Multimodal Model for Predicting Conversational Feedbacks A Multimodal Model for Predicting Conversational Feedbacks. (2021). https://doi.org/10.1007/978-3-030-83527-9_46Ãŕ

[6] Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A Shallow Model of Backchannel Continuers in Spoken Dialogue. www.hcrc.ed.ac.uk/

[7] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[8] Pino Cutrone. 2005. A case study examining backchannels in conversations between Japanese–British dyads. (2005).

[9] Rory T Devine, Naomi White, Rosie Ensor, and Claire Hughes. 2016. Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental psychology* 52, 5 (2016), 758.

[10] Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2020. Quantifying the interplay of conversational devices in building mutual understanding. (2020).

[11] Allen T Dittmann. 1972. Developmental factors in conversational behavior. *Journal of Communication* 22, 4 (1972), 404–423.

[12] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.

[13] Gabriel Doyle, Dan Yurovsky, and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in twitter conversations. *25th International World Wide Web Conference, WWW 2016*, 637–648. https://doi.org/10.1145/2872427.2883091

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 1459–1462. https://doi.org/10.1145/1873951.1874246

[15] Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 199–206.

[16] Dmitriy Genzel and Eugene Charniak. 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*. 65–72.

[17] Mononito Goswami, Minkush Manuja, and Maitree Leekha. 2020. Towards Social Engaging Peer Learning: Predicting Backchanneling and Disengagement in Children. (7 2020). http://arxiv.org/abs/2007.11346

[18] Agustín Gravano and Julia Hirschberg. 2009. Backchannel-Inviting Cues in Task-Oriented Dialogue.

[19] Beata Grzyb, Stefan L Frank, and Gabriella Vigliocco. 2022. Communicative efficiency in multimodal language. (2022).

[20] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7, 1 (2017), 411–420.

[21] Vidit Jain and Maitree Leekha. 2021. Exploring semi-supervised learning for predicting listener backchannels. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445449

[22] Rajni Jindal, Maitree Leekha, Minkush Manuja, and Mononito Goswami. 2020. What makes a better companion? towards social & engaging peer learning. In *ECAI 2020*. IOS Press, 482–489.

[23] Judith R. Johnston. 1988. Acquisition of Back Channel Listener Responses to Adequate Messages. *Discourse Processes* 11 (1988), 319–335. Issue 3. https://doi.org/10.1080/01638538809544706

[24] Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45 (2017), 326–347.

[25] Michael McCarthy. 2002. Good listenership made plain. *Using corpora to explore linguistic variation* 9 (2002), 49.

[26] Louis Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20 (1 2010), 70–84. Issue 1. https://doi.org/10.1007/s10458-009-9092-y

[27] Michael Murray, Nick Walker, Amal Nanavati, Patricia Alves-Oliveira, Nikita Filippov, Allison Sauppe, Bilge Mutlu, and Maya Cakmak. 2022. Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. In *Conference on Robot Learning*. PMLR, 513–525.

[28] Elias Chaibub Neto, Abhishek Pratap, Thanneer M. Perumal, Meghasyam Tummalacherla, Phil Snyder, Brian M. Bot, Andrew D. Trister, Stephen H. Friend, Lara Mangravite, and Larsson Omberg. 2019. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *npj Digital Medicine* 2 (12 2019). Issue 1. https://doi.org/10.1038/s41746-019-0178-x

[29] Hiroaki Noguchi and Yasuharu Den. 1998. Prosody-based detection of the context of backchannel responses. In *Fifth International Conference on Spoken Language Processing*.

[30] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. *ACM/IEEE International Conference on Human-Robot Interaction* Part F127194, 100–108. https://doi.org/10.1145/2909824.3020245

[31] Carole Peterson. 1990. The who, when and where of early narratives. *Journal of child language* 17, 2 (1990), 433–455.

[32] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings.. In *Interspeech*. 879–883.

[33] Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology* 16 (10 2020). Issue 10. https://doi.org/10.1371/journal.pcbi.1008228

[34] Marilyn Shatz and Rochel Gelman. 1973. The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the society for research in child development* (1973), 1–38.

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

[36] Khiet Truong, Dirk Heylen, Khiet P Truong, and Ronald Poppe. 2010. A rule-based backchannel prediction model using pitch and pause information Tasty Bits and Bytes View project Interaction for Universal Access: Socially Intelligent Agents in Serious Gaming Environments View project A rule-based backchannel prediction model using pitch and pause information. https://www.researchgate.net/publication/221489093

[37] Bekir Berker Türker, Engin Erzin, Yücel Yemez, and T Metin Sezgin. 2018. Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions.. In *Interspeech*. 1741–1745.

[38] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. , 1177-1207 pages. www.elsevier.nl/locate/pragma

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[40] Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition* 170 (1 2018), 147–163. https://doi.org/10.1016/j.cognition.2017.09.018

[41] Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*. 567–578.

[42] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).