ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200129

What Makes a Better Companion? Towards Social & Engaging Peer Learning

Rajni Jindal[*†] and **Maitree Leekha**[*†] and **Minkush Manuja**[*†] and **Mononito Goswami**[*†]

Abstract. Peer learning companions such as interactive tablets and social robots have shown great promise in supporting language development in young children. However, studies have shown that the perceived credibility of a robot as an educator and peer companion is contingent on how socially it behaves. We specifically focus on two roles of a peer learning companion- as an engaging storyteller and active listener. To this end, we develop models to predict whether the listener will lose attention (Listener Disengagement Prediction, LDP) and whether the robot should generate listener backchannels with high probability (Backchanneling Extent Prediction, BEP) during a specific time window. We formulate LDP and BEP as Time Series Classification problems and through extensive evaluation in multiple experimental settings, demonstrate our models' promising results. Inspired by prior work, we also investigate socio-demographic and developmental features, which may give rise to variations in children's backchanneling responses. Moreover, we examine critical features responsible for the predictive utility of our models using Permutation Feature Importance and Partial Dependency Plots. Our findings suggest that features such as pupil dilation, blink rate, acceleration of head, gaze direction, and some facial action units which have not been considered in prior work, are in fact, critical in predicting backchanneling extent and listener disengagement.

1 Introduction

Studies such as [31] have shown that children who are encouraged to spend more time in narrative conversations display a significant growth of vocabulary and an overall increase in narrative skill. For young children, effective narration can be used to promote school readiness and is an essential prerequisite to successful communication [9]. Conversational activities such as storytelling, that hone narrative skills, thus play a pivotal role in children's early language development. However, delivering engaging narratives requires a successful back-and-forth process involving speaker cues and listener backchannels (BC). Listener backchannels are non-verbal signals which indicate that the communication is working, and the speaker must continue speaking [38].

Peer learning companions such as interactive tablets, computer applications, and social robots have the potential to support early language learning in children. Through adequate support and scaffolding, peer tutors steer the learning process to remain in the "Zone of Proximal Development". Topping and Ehly [37] provided a theoretical model to explain the cognitive benefits of peer tutoring and suggested that peer tutors act as co-learners and minimize their companion's frustration as a result of challenges and impasses. Our work is a first step towards developing engaging personal learning companions, which can promote early language development in children by leveraging the benefits of technology. Prior research has shown that young children may not only consider robots as emotional and trustworthy social beings [21] but also willingly learn new information from them [26]. Therefore, given the practical benefits of peer tutoring and the ability of robots to act as reliable social beings, recent studies have attempted to harness the potential of robots as peer learning companions to foster the development of early language skills in children [29]. However, the credibility of robots as peer learning companions depends on how socially-contingent they behave [4].

In this paper, we focus on two roles of a social peer learning companion; as an engaging storyteller and active listener. An engaging storyteller must be able to predict if the listener will lose attention, to initiate an action that precludes listener disengagement. On the other hand, a listener must also actively communicate with the storyteller through backchannels. However, predicting when a listener should generate a backchannel response is a hard problem. To this end, we develop two models to predict (*i*) whether the listener will lose attention (Listener Disengagement Prediction, 'LDP' model), and (*ii*) the extent to which a listener should generate backchannel responses in the next few (3) seconds (Backchanneling Extent Prediction, 'BEP' model).

Most research papers in the past have utilized a limited set of features to predict listener engagement and backchanneling opportunities, *i.e.* predicting the exact time at which a listener should backchannel. To model listener attention, Lee *et al.* [23] used a combination of manually annotated categorical features like pitch, energy, gaze, *etc*. Furthermore, in their work [29], Park *et al.* analyzed the speaker cues which were useful in eliciting positive responses from listeners, and found that categorical prosodic cues like children's speech being too wordy, their pitch, energy, and long pauses, were useful in predicting backchanneling opportunities. However, we believe that these manually annotated categorical features limit our understanding of backchanneling behaviors and engagement states, especially when analyzing the narratives of young children. In this regard, we investigate whether incorporating additional features such as pupil dilation, blink rate and acceleration of head automatically, and considering their dynamics improves the prediction capability of our LDP and BEP models. Although we only predict the extent of backchanneling ('high' or 'low') in the next time window instead of the specific time at which listeners should backchannel, we believe that our model can be easily deployed as an opportunity prediction model due to the small size (3 seconds) of the windows fed into our

[*] All authors contributed equally, and wish that they be regarded as joint First Authors.
[†] Department of Computer Science & Engineering, Delhi Technological University, India, Email: rajnijindal@dce.ac.in, {maitreeleekha, minkushmanuja, mononito}_bt2k16@dtu.ac.in

models. We also explore whether socio-demographic factors such as gender, mother's highest education, and household income of the storyteller and listener influence the extent of backchanneling generated by the listener. To further understand our findings, we also examine the most critical features responsible for the predictive capability of our models using Permutation Feature Importance [1] and Partial Dependency Plots (PDPs) [16].

The rest of the paper is organized as follows. The next section explores prior literature and its relation to the present work. Section 3 discusses the dataset used in this paper, feature engineering, and the LDP and BEP models. Section 4 describes the experimental settings and the results of our detailed experiments. Finally, Section 5 concludes our work and discusses avenues of future work.

2 Relation to Prior Work

Effective communication is a collective activity of the first order. Both speakers and listeners must transcend beyond merely issuing utterances and, listening and understanding, respectively. Instead, they must coordinate on the content of the conversation and make sure that they have established a mutual understanding on the subject of the discussion. The speaker and listener must also coordinate the process, *i.e.* the speaker must talk only when the listener appears to be hearing and trying to understand, and the listener should communicate the same to the speaker [8].

Listeners communicate through backchannel responses in the form of gestures such as gaze locking, nodding, and short verbal expressions (*yeah, oh*). These backchanneling responses serve multiple cognitive functions, such as indicating the state of engagement, understanding, and sentence completion [8]. While there has been substantial research on adult speaking and listening, there is surprisingly little work investigating the same for young children, especially in the context of dyadic interactions [23]. Most prior work has focused on adult-child pairs and has demonstrated the influence of age on backchanneling responses [19, 12]. Lee *et al.* [23] were amongst the first to identify attention-related listener backchannels, and show that both the listener's and speaker's behaviors must be taken into account to best infer the attentive state of the listener. Drawing inspiration from their work, we use features from both the speaker and listener to predict the attentive state of the latter.

The timing of backchanneling responses presents another significant technical challenge. Most of the previous approaches have been able to successfully detect backchanneling opportunities using prosodic features such energy and pitch [25, 28]. Morency *et al.* [25] presented a real-time backchanneling prediction model using vocal prosodic features and speaker gaze. They demonstrated how sequential probabilistic models such as Hidden Markov Models (HMMs) and Conditional Random Fields can automatically learn from human-to-human interactions to predict listener backchannels. Poppe *et al.* [33] also utilized features from the speaker's speech and gaze, and some rule-based strategies to predict the placement of backchannels. However, most of these models learned from adult behaviour and were trained on adult voices. To this end, Park *et al.* [29] pioneered the development of a backchanneling opportunity prediction model for young children. They identified the speaker cues useful in eliciting backchanneling responses from the listener, and using these features predicted the backchanneling opportunities. Specifically, they found four prosodic cues of the speaker useful- pitch, long pauses in between the speech, energy, and the speech being too verbose (wordy).

Our work differs from prior literature in the following ways. First, unlike most work on backchanneling and listener engagement using categorical vocal features, we use rich, continuous-valued prosodic features like pitch, Mel-Frequency Cepstral Coefficients (MFCC) and RMS Energy. We also incorporate features such as Facial Action Units (FAUs), pupil dilation, blink rate, velocity and acceleration of the head and eye gaze to predict listener disengagement and the extent of listener's backchannel responses. Facial action units [34] have been widely used to detect prototypical facial expressions and infer emotions. Moreover, some studies [17, 13, 3] have also shown that changes in gross body movements such as the velocity of head characterize cognitive states and depression severity. Eye aspect ratio [18] and the size of the pupil [22] have also been shown to be reliable indicators of the alertness, as well as the ability of the speakers to elicit backchannel responses, respectively. Furthermore, most of the approaches so far have utilized features aggregated across time windows to predict backchanneling opportunities or infer the attentive state of the listener [23]. However, we hypothesize that the dynamics of features present useful information that can be utilized to better predict the attentive state of the listener and backchanneling responses. Thus, we also experiment with ResNet [15], which can learn useful patterns from multivariate time series, to predict listener disengagement and extent of backchanneling responses.

3 Methodology

3.1 Dataset

Datasets capturing social interaction are pivotal to understanding social interactions and subsequently designing human-like social technologies. However, most existing datasets such as ALICO [24] and MultiLis [10] have emerged from attempts at studying interactions between adults. Social interactions of children have been analysed through child–adult pairs only. Thus, while prior literature has closely studied adult-adult and adult-child interactions, there has been little work on analysing the dynamics of child–child interactions.

Since our work aims to develop peer learning companions to support the early language development in children, we use the P2PSTORY dataset [35], which is one of the first attempts at investigating the social and emotional behaviors of children through storytelling exercises amongst peers. For their study, the authors recruited eighteen kindergarten participants from diverse cultural backgrounds, with an average age of 5.22 years and varying levels of social and emotional development. In the storytelling exercises, each child participated in at least three rounds of storytelling with different partners, and building stories with the help of proctors based on text-less storybooks. In a particular *dyad session*, a pair of children took turns to narrate stories to their partner, with each turn generating a *storytelling episode*. The dataset comprises of 58 storytelling episodes with an average length of 1 minute 17 seconds.

For each episode, three time-synchronized cameras captured the front-views of each participant and the bird's eye view of the dyad. In addition to these videos, the authors also provided high-quality audio recordings for the sessions. Using the video recordings and ELAN [7] (a video annotation software), the dataset creators also annotated several nonverbal behavioral features[‡]: gaze (0.89), posture (0.40), nod (0.89), eyebrow movement (0.51), mouth (0.34), utterances (0.81), voicing (0.83), on/off task (0.05), attentive state (0.45). To analyze the influence of socio-demographic and developmental features on the acquisition of speaker cues and listener responses, the authors also released the results of the Ages and Stages Question-

[‡]Inter-rater reliability was measured using Fleiss' kappa κ and has been reported in parenthesis.

naire (ASQ) for all the participants. ASQ is a standardized measure to evaluate the social and emotional development of children. The interested reader may refer their paper [35] for more details on the data and study procedure.

3.2 Feature Engineering

As discussed in Section 2, in addition to the annotations provided in the dataset, we also extracted several features along the visual and audio channels. We used OpenFace [2] at a sampling frequency of 30 Hz to extract Facial Action Units (*FAUs*), head and eye gaze orientations. We used the eye-landmarks extracted using OpenFace as shown in Figure 2(ii), to determine *blink rate* and the *pupil dilation*. We computed the derivatives of the visual features with the exception of FAUs, to obtain velocity and acceleration of eye gaze and head. To this end, we used NumPy's[§] gradient function which approximates the gradient of an array using second-order accurate central differences in the interior points and second-order accurate one sides in the endpoints. For each storytelling episode, we also used OpenSmile [14] to extract several vocal prosodic features from the audio recordings of the speaker at a frequency of 30 Hz. Jiang *et. al* [20] had established the utility of both audio and visual features in increasing emotion recognition accuracy. Inspired by their work, we too extracted the fundamental frequency, Mel-Frequency cepstral coefficients (MFCC) together with their first and second order derivatives, and Root-mean-square energy for our prediction tasks. Tables 1 and 2 summarize all the visual and prosodic features used in our study.

3.3 Listener Disengagement Prediction

A peer learning companion must be an engaging speaker. To this end, we attempt to predict the attentive state of the listener after a specific time window. Such a capability would be very useful for artificial speakers, as they may be able to initiate an engaging action or utterance to preclude listener disengagement.

In this work, we model the listener's attentive state as a time series classification problem. We first divide all the storytelling episodes into non-overlapping windows of 3 seconds each. The choice of window size is inspired by the recent study carried out by Park *et. al* [30]. Moreover, a 3 second window having 90 time steps[¶] is short enough to be stationary (long time series capturing behavioural data tend to be non-stationary [27]), but still has enough information to allow ResNet to model the dynamics of its features. In the classification task, for each window we use the prosodic cues of the speaker, and visual and behavioral cues of both the partners, for the first 89 time steps to predict the listener's attentive state at the 90^{th} time step. However, the OpenFace and OpenSmile features were extracted at a higher sampling frequency (30 Hz) than the behavioural annotations (5 Hz) available with the dataset. To resolve this mismatch, we aggregated the behavioural annotations into window-level features. Thus, each window had two sets of features: dynamic visual and prosodic features in the form of a multivariate time series with 89 time steps, and a vector of aggregated behavioural features. In order to aggregate behavioural features which were categorical in nature, we introduced dummy binary variables and reported the ratio of time steps that each of variables were *True*. For example, at each time step the speakers and listeners can either *gaze* towards their *partners*, the *picture* (storyboard) or *away*. Therefore, we introduced three dummy variables as a substitute for *gaze*: *gaze_partner*, *gaze_picture* and *gaze_away*. We then computed the ratio of time steps that each of these dummy binary variables were *True*, to the total number of time steps in a window *i.e.* 89 time steps. We did not up-sample the behavioural features and use them as time series features instead, because we found that the behavioural features rarely changed values in a 3 second time window, and therefore the dynamics of behavioural features must contain little additional information. Moreover, by reducing the number of time series features we also reduced the training and testing time of our models.

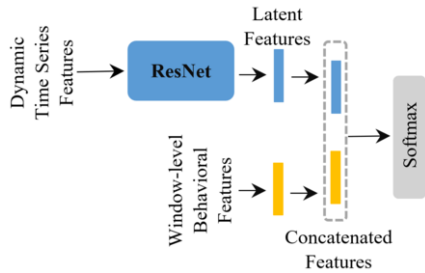


Figure 1. ResNet model for LDP using dynamic time series and window-level behavioral features.

We mathematically formulate the LDP task as follows. Let, $\mathtt{W}_i$ be the $\mathtt{i}^{\mathtt{th}}$ window, and $\mathtt{T}_j$ represent the time series visual and prosodic features of the listener and the speaker for the $\mathtt{j}^{\mathtt{th}}$ time step in the window. Therefore, $[\mathtt{T}_1, \mathtt{T}_2, ..., \mathtt{T}_{89}]_i$ represents the dynamic features for the first 89 time steps in $\mathtt{W}_i$. Also, let $\mathcal{B}_i$ be the combined window-level representation of the behavioral features for $\mathtt{W}_i$. Then, for the speaker to predict the attentive state, $\mathcal{A}_i$, of the listener at the $90^{\mathtt{th}}$ time step of the window $\mathtt{W}_i$, we need to learn the following mapping function $\mathcal{F}$,

$$\mathcal{F}([\mathtt{T}_1, \mathtt{T}_2, ..., \mathtt{T}_{89}]_i, \mathcal{B}_i) \mapsto \mathcal{A}_i \tag{1}$$

We experimented with two learning algorithms that have been extensively used for time series classification: Random Forests and ResNet [15]. To train Random Forests, we aggregated the OpenFace features by taking their arithmetic mean and combined them with the window-level behavioral features. To model the dynamic visual and prosodic features, we also used ResNet, which is a state-of-the-art deep learning-based time series classification model, and has been shown to perform comparably to HIVE-COTE [15]. In addition to these dynamic features, we also use the window-level behavioral features, by concatenating them with the latent feature representation of the dynamic features learned by ResNet, and then feeding them all to the softmax layer, as shown in Figure 1.

3.4 Backchanneling Extent Prediction

Backchannel responses are essential to effective communication as they serve as feedback to speakers, enabling them to recognize the extent to which their listeners understand them [11]. In an intervention study, Carole *et al.* [32] had found that backchanneling while talking to children had a signicant impact on their overall narrative skills. Therefore, an active artificial listener must learn to backchannel appropriately. The timing of backchannel response is a challenging technical problem and has been solved in prior work [30]. Our BEP model differs from prior literature since it does not specify a time to produce listener backchannels. Instead, using our model we take a deeper look at the speaker cues which prompt *high* and *low* levels of backchanneling from listeners.

Like listener disengagement prediction, we use time series classification to predict backchanneling opportunities using 3 second windows and features sampled at a frequency of 30 Hz. Using the speaker's vocal prosodic and visual features extracted for a window

[§]https://numpy.org/

[¶]Features are sampled every 30 ms (Section 3.2).

Features	Description	Derivation from OpenFace features
FAUs	Indicate the presence or absence of 18 Facial Action Units	AU01_c, AU02_c, AU04_c, AU05_c, AU06_c, AU07_c, AU09_c, AU10_c, AU12_c, AU14_c, AU15_c, AU17_c, AU20_c, AU23_c, AU25_c, AU26_c, AU28_c, AU45_c
gaze_vel	Velocity of eye gaze	$\sqrt{(\texttt{gaze\_angle\_x})'^2 + (\texttt{gaze\_angle\_y})'^2 + (\texttt{gaze\_angle\_z})'^2}$
gaze_acc	Acceleration of eye gaze	$\sqrt{(\texttt{gaze\_angle\_x})''^2 + (\texttt{gaze\_angle\_y})''^2 + (\texttt{gaze\_angle\_z})''^2}$
head_vel_T	Translational velocity of head	$\sqrt{(\texttt{pose\_Tx})'^2 + (\texttt{pose\_Ty})'^2 + (\texttt{pose\_Tz})'^2}$
head_vel_R	Rotational velocity of head	$\sqrt{(\texttt{pose\_Rx})'^2 + (\texttt{pose\_Ry})'^2 + (\texttt{pose\_Rz})'^2}$
head_acc_T	Translational acceleration of head	$\sqrt{(\texttt{pose\_Tx})''^2 + (\texttt{pose\_Ty})''^2 + (\texttt{pose\_Tz})''^2}$
head_acc_R	Rotational acceleration of head	$\sqrt{(\texttt{pose\_Rx})''^2 + (\texttt{pose\_Ry})''^2 + (\texttt{pose\_Rz})''^2}$
blink_rate	First order differential of Eye Aspect Ration (averaged for both the eyes) [36]	$[(\|\|e_{18} - e_{10}\|\| + \|\|e_{16} - e_{12}\|\|)/(2 * \|\|e_{14} - e_8\|\|)]'$
pupil_dilation	Size of pupil (averaged for both the eyes)	$(\|\|e_{25} - e_{21}\|\| + \|\|e_{27} - e_{23}\|\|)/(\|\|e_{14} - e_8\|\| + \|\|e_{17} - e_{11}\|\|)$

Table 1. Visual features extracted for both the speaker and listener. X is the returned by OpenFace. $(\texttt{X})'$ and $(\texttt{X})''$ are its first and second order derivatives.

Feature	Description
F0	The fundamental frequency computed from the Cepstrum
$\texttt{mfcc}', \texttt{mfcc}''$	First and Second order derivatives of Mel-Frequency Cepstral Coefficients 1-12
$\texttt{pcm\_RMSenergy}', \texttt{pcm\_RMSenergy}''$	First and Second order derivatives of Root-mean-square signal frame energy

Table 2. Prosodic features for the speaker. X is the feature returned by OpenSmile. $(\texttt{X})'$ and $(\texttt{X})''$ are its first and second order derivatives.

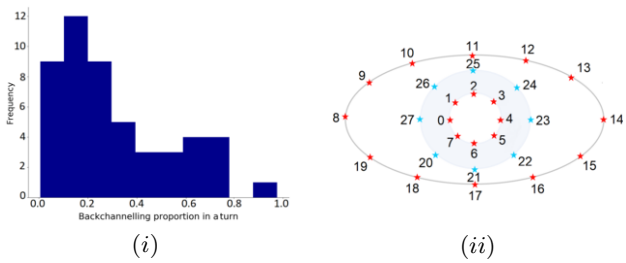


Figure 2. (i) Distribution of backchanneling proportion across sessions (ii) Eye landmarks extracted using OpenFace and used to calculate Eye Aspect Ratio and Pupil Dilation Rate.

$\texttt{W}_i$, we predict the extent of listener's backchannel response in the window $\texttt{W}_{i+1}$. Inspired by the research carried out by Dennis *et. al* [11], we consider both verbal (listener utterances such as *"hmm"*, *"ooh"*, etc.) and non-verbal (smile, nod, onset of partner gaze, lean towards, raising brows) backchannels in our model. To classify a window as being indicative of '*high*' or '*low*' backchanneling, we first find the proportion (p) of time steps in the window, where the listener generates at least one backchannel response. Then to convert these continuous proportions associated with the windows to a binary label (*high backchanneling* or *low backchanneling*), we use a threshold τ such that if $p > \tau$, the window indicates *high backchanneling*. Instead, if $p \leq \tau$, we label the window as *low backchanneling*. The threshold τ controls the length of time ($\tau \times 3$ seconds) that qualifies as *high backchanneling*. In this work, we experiment with $\tau = 0.25$ and $\tau = 0.5$. A higher threshold τ includes backchannels that last longer, whereas a lower value includes subtle as well as long-lasting ones.

We now formally frame the problem of backchanneling extent prediction. Let $\texttt{W}_i$ be any 3 second window, and let $[\texttt{T}_1, \texttt{T}_2, ..., \texttt{T}_{90}]_i$ represent the dynamic prosodic and visual features of the speaker for $\texttt{W}_i$. Furthermore, let $\mathcal{P}_{i+1} \in \{$*high backchanneling*, *low backchanneling*$\}$ be the target variable capturing the extent of backchanneling in the next window, $\texttt{W}_{i+1}$. Therefore, the BEP task involves learning the following function $\mathcal{G}$,

$$\mathcal{G}([\texttt{T}_1, \texttt{T}_2, ..., \texttt{T}_{90}]_i) \mapsto \mathcal{P}_{i+1} \tag{2}$$

Like listener disengagement prediction, we used both Random Forests and Resnet to predict the extent of backchanneling. The window-level arithmetic mean of the speaker's prosodic and visual features were used as featurization for Random Forests, whereas Resnet used the dynamic time series features.

3.5 Socio-demographic factors influencing backchanneling

Prior studies [19, 12] have shown that young children demonstrate significant variation in terms of their backchanneling responses. To shed more light on the factors giving rise to these variations, we examine how socio-demographic features such as household-income and mother's education influence the extent of backchanneling. We quantify the extent of backchanneling in a storytelling session in terms of the *backchanneling proportion* or the ratio of time steps which were annotated as having some form of backchanneling to the total number of annotated time steps in a session. The distribution of backchanneling ratio across different sessions is illustrated in Figure $2(i)$.

We hypothesize that the socio-demographic features shown in Table 6 lead to significant differences in the distribution of backchanneling proportions in different storytelling sessions. To test our hypothesis, we conducted two-sample Kolmogorov-Smirnov (K-S) tests which compare the empirical cumulative distributions (ECDFs) of two samples to see whether they differ significantly. The null hypothesis of K-S test is that the two samples are drawn from the same underlying distribution. A notable exclusion from the set of socio-demographic features tested for difference, is the age of the children. We did not include age as a factor since most of the children were of the same age, and most prior studies [19, 12] have already studied its effect on the extent of backchanneling responses. Furthermore, to analyse the influence of developmental factors in detail, we carried out K-S tests on the responses to some questions of the ASQ questionnaire too (Table 6). The two samples for the K-S tests were

either two values of the features or the two most frequently occurring values. For example, to see whether the storyteller's gender has any influence on the backchanneling proportion, the two samples belonged to `male` and `female` storytellers, respectively. However, features such as Mother's highest educational qualification had more than two values such as `Graduate or professional training`, `College Graduate`, `Some high school` etc., and therefore we used the two most frequently occurring qualifications as our two samples[‖]. In order to test the impact of household income on backchanneling responses, we divided the household incomes into categories: high-income (Over $100,000) and low-income ($30,000 to $75,000), respectively. Similarly, we also categorized children as having `low` or `high` ASQ scores based on multiple thresholds (20, 25 and 30), such that children having ASQ scores less than the threshold have `low` scores and the other children have `high` scores. Consequently, children having `high` and `low` scores belong to the two samples. We also carried out a randomization test (1000 runs) and computed the K-S statistics (*D*) by randomly permuting the feature values. Our findings are discussed in detail in Section 4.3.

4 Results

To evaluate our models we conducted three kinds of experiments given the hierarchical nature of our data: two children (subjects) form a dyad and participate in two turns of storytelling (episodes), where each child narrates a story once, and listens the other time. The first type of experiment ("Random stratified") makes a stratified 5-fold train and test split across all the windows based on the predicted labels. In the second kind of experiment ("Leave-One-Episode-Out" or LOEO) we set a particular episode and all the windows associated with it aside as the test set, and used remaining data to train our models. In the final ("Leave-One-Subject-Out" or LOSO) experiment, we set aside all the windows corresponding to a particular child as the test set, and use the remaining data as the train set.

We expect the performance of our models to degrade from stratified random sampling to the LOSO experiments due to the extent of information sharing between the training and testing sets. In random stratified experiments, both the train and test sets comprise of information about a subject and episodes. In contrast, in the LOEO experiments the model might have seen the subject in the training set, but it does not have any information about the episode it is being validated on. Thus, LOEO experiments represent a "warm-start" situation, since the model has some knowledge about the subjects, but has no information about the current episode. The LOSO experiments represent *cold-start* since the model does not have any information about the subject (the subject may either be the speaker or listener).

Besides developing effective classifiers, it is also important to analyze critical features which influence the classification models. To this end, we examined the importance of features fed into the Random Forests using two measures: Mean Decrease Impurity and Permutation Feature Importance.

Random Forests use the gini criterion to split on a variable[**]. At every step of the tree construction, one of the n variables is chosen to form a split conditioned on its values, which results in a decrease in node impurity (gini). Consequently, mean decrease in impurity (MDI) computes the significance of a variable by measuring the total decrease in gini as a result of splitting on it, weighted by the number of samples and averaged over all the trees. Therefore, important features have high values of MDI [6]. However, MDI tends to magnify the importance of continuous and categorical variables with high cardinality. Therefore, we supplement our analysis using permutation feature importance [5] or mean decrease in accuracy (MDA) which is not only impartial to both continuous and high-cardinality categorical variables, but also tells us whether Random Forests have over fit. They may have over fit to the training data, if 'important' features returned by MDI and MDA are significantly different.

Permutation feature importance relies on the principle that if a feature is important, the model relies on its values for the prediction, and thus randomly shuffling its values must decrease the model's accuracy (or increase the error). On the other hand, values (or samples) of an unimportant feature should have negligible effect on the model's accuracy. Permutation importance for a feature $\mathcal{X}$ is calculated by repeatedly permuting (s times) its values and averaging the difference in the error computed before and after the permutation, over all the trees. MDA can be computed on both the training and testing sets, which have as yet undiscovered yet slightly different implications. However, in this study, MDA is computed on the test set to rule out the possibility of classifying unimportant features as important ones. Overfit Random Forests may return spurious features as important if MDA is computed on the training set.

In order to analyse the relationship of important features with the outcome variable, we generated partial dependency plots for the two most important features for all Random Forest models. Partial Dependency plots (PDPs) demonstrate the marginal effect of a subset of features on the outcome of a machine learning model, and are useful in examining whether the relationship between the target and the feature is linear, monotonic or complex [16].

4.1 Predicting Listener Disengagement

Detailed results of predicting the attentive state of the listener using Random Forests and ResNet are shown in Table $3(i)$. In the table, we only report metrics for the "not-listening" (negative) class because we are interested in predicting disengagement accurately. Moreover, we had very few samples for "not-listening" in comparison to the "listening" class and therefore even weighted metrics would have been biased in favour of "listening". We found that both our models had limited predictive utility due to the fuzzy nature of the target variable. Even trained annotators only had moderate agreement (Fleiss' $\kappa = 0.45$) while identifying the attentive states. Nevertheless, we noticed an interesting phenomenon in Table $3(i)$; ResNet which takes into account the dynamics of OpenFace time series features performed better than Random Forests trained on aggregated features, for the most part. Since it is more important to predict "not listening" accurately, we must focus on improving the recall of the "not listening" (negative) class. In all the experiments, ResNet consistently has a better recall than Random Forests, and therefore it classifies "not-listening" more accurately. This is desirable for engaging peer learning companions as they will be able to better predict when the listener is going to lose attention.

To identify features essential for predicting the listener's engagement state, we calculated feature importance using MDI and MDA for the experiments (random, LOEO and LOSO). Although all the experiments identified the same set of features as important, we found some minor variations in their rankings across the experiments. In the remainder of the paper, we only report feature importance results for LOSO experiments since they capture the most general view of the problem at hand. Moreover, important features derived from LOSO experiments do not over fit on the peculiarities of individual subjects.

[‖] We observed that most of the mothers were either `College Graduates` (44%) or `Professionally trained graduates` (38%).

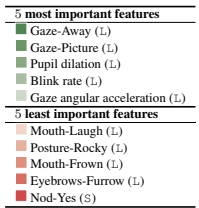
[**] Note that we used the term variable and feature interchangeably.

Table 3(ii) illustrates the 5 most important (■), and the 5 least important (■) features derived using MDI. We found that the listener's *gaze-away* and *gaze-picture* had the highest importance values for predicting listener disengagement. This is probably because children gazing away may be perceived as being distracted and not paying attention [23]. Moreover, features derived using OpenFace such as *pupil dilation*, *blink rate* and *gaze angular acceleration* also proved to be important. To the best of our knowledge, these features have not been used to analyze peer-to-peer social interactions in the literature so far. The MDI values also show that most of the behavioral features such as *eyebrows*, *nod* and *posture* are less useful in predicting listener disengagement. This may be because of noise in the features[††].

Figure 4(i) shows the box and whisker plot of the 10 most important features for predicting the listener disengagement obtained using MDA ($s = 50$). Both MDI and MDA agree on the most important features for predicting listener's loss of attention. Figures 3(ii) and 3(iii) illustrate the marginal impact of listener's *gaze-away* and *gaze-picture* on his attentive state using PDPs. We can clearly see that listeners are considered to be more inattentive when they look (gaze) away from the storyteller. In contrast, gazing at the picture (storyboard) has an incremental effect on the listener's attention.

Model	P	R	F1	AUC
Random split				
Random Forest	**0.75**	0.58	0.65	**0.91**
ResNet	0.70	**0.67**	**0.71**	0.82
Leave one episode out (LOEO)				
Random Forest	0.56	0.75	0.62	0.85
ResNet	**0.66**	**0.78**	**0.68**	**0.91**
Leave one subject out (LOSO)				
Random Forest	0.63	0.78	0.67	0.85
ResNet	**0.67**	**0.79**	**0.70**	**0.89**

(i)

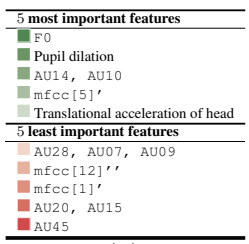


(ii)

Table 3. (i) Listener Disengagement Prediction (LDP): we report the P, R, F1 and AUC for the "not listening" class. (ii) Mean Decrease in Impurity: 5 most and least important features for Random Forests for LDP (L: Listener, S: Speaker.)

τ	Model	P	R	F1	AUC
	Random split				
0.25	**Random Forest**	**0.91**	0.48	0.63	0.86
	ResNet	0.68	**0.92**	**0.71**	**0.88**
0.50	**Random Forest**	**0.68**	**0.99**	**0.80**	**0.86**
	ResNet	0.65	0.79	0.70	0.83
	Leave one episode out (LOEO)				
0.25	**Random Forest**	**0.90**	0.49	0.60	0.86
	ResNet	0.69	**0.93**	**0.73**	**0.89**
0.50	**Random Forest**	**0.70**	**0.97**	**0.80**	**0.86**
	ResNet	0.68	0.78	0.70	0.77
	Leave one subject out (LOSO)				
0.25	**Random Forest**	**0.90**	0.57	0.68	0.87
	ResNet	0.71	**0.91**	**0.73**	**0.90**
0.50	**Random Forest**	**0.66**	**0.98**	**0.78**	**0.86**
	ResNet	**0.66**	0.85	0.76	0.78

(i)



(ii)

Table 4. (i) Backchanelling Extent Prediction (BEP): we report metrics for the "high backchanneling" class. (ii) Mean Decrease in Impurity: 5 most and least important features for Random Forests for BEP.

4.2 Predicting the Extent of Backchanneling

We trained ResNet and Random Forests to predict the extent of backchanneling in the next window w_{i+1} for two values of τ, 0.25 and 0.5. For the same reasons as LDP, we only report metrics for the "high backchanneling" or positive class. We gained several interesting insights from the BEP results summarized in Table 4(i). First, the performance of ResNet in terms of F1, remains roughly the same for all experimental settings including $\tau = 0.25$ and 0.50. Furthermore, Random Forests is able to predict the extent of backchanneling better than the ResNet for $\tau = 0.50$ *i.e.,* for long-lasting backchannels. However, when we introduce subtle and shorter backchannels in the training data by setting $\tau = 0.25$, the performance of Random Forests drops, which is evident from the sharp decrease in recall. When we reduce the threshold τ from 0.5 to 0.25, the recall drops from an average of 0.98 to 0.51, computed across all the experimental settings (random, LOEO and LOSO). A plausible reason for the sharp drop for $\tau = 0.25$, is the loss of information due to aggregation which causes Random Forests to be unable to predict short and subtle backchannel responses. However, feature aggregation does not impact the performance of our model in case of $\tau = 0.50$, possibly because it is able to clearly distinguish "high backchannelling" from "low backchanneling". It must be noted that the difference between "high" and "low" backchanneling becomes fuzzier as the threshold drops from $\tau = 0.5$ to 0.25. In summary, we believe that ResNet predicts short and subtle backchannels better than Random Forest because it uses dynamic time series features.

Table 4(ii) summarizes the 5 most (■) and least important (■) features used for predicting the extent of backchanneling, computed using MDI for $\tau = 0.50$. As shown in prior work, the pitch (F0) of the speaker is one of the most important factors influencing listener backchannels. Besides, we also observe that the speaker's *pupil dilation*, AU10 (*upper lip raiser*), AU14 (*dimpler*), and *translational head acceleration* also help in eliciting backchannel responses from the listener. The derivatives of mel-frequency cepstral coefficients (mfcc'), which represent the envelop of time power spectrum of speech signals are also of high importance. In contrast, features such as AU20 (*lip stretcher*), AU15 (*lip corner depressor*) are relatively less important.

Figure 4 illustrates the 10 most important features to predict backchanneling using 50 rounds of MDA ($s = 50$). As per our expectations the speaker's F0 (pitch), *pupil dilation* and AU14 (*dimpler*) are the most important factors according to MDA. To further analyze the combined influence of F0 and *pupil dilation* on the extent of backchanneling, we plotted a 3-dimensional PDP shown in Figure 3(i). It is evident that both these features have a positive impact in seeking backchannel responses. It is interesting to note that around the range of $0.20 - 0.22$, the speaker's *pupil dilation* causes a sharp increase in the listener's backchanneling response. In summary, our results emphasize the importance of features such as the speaker's pupil dilation and previously used features like the speaker's pitch (F0) in predicting backchannel reponses from young listeners.

4.3 Socio-Demographic Analysis

The results of the K-S and randomization tests are summarized in Table 6. It can be clearly seen that, amongst all the socio-demographic and developmental factors considered, whether listeners look at their parents while talking to them ($\mathcal{D} = 0.569$) and the listeners' mothers' education ($\mathcal{D} = 0.591$) influence the distribution of BC proportion the most. In order to examine the direction of difference, we plotted Empirical Cumulative Distribution Functions (ECDFs) for each of the factors which yielded significant differences. For example, on analysing the ECDFs in Figure 2(i), we concluded that for a given BC proportion $\mathcal{P}$, it is much more likely to observe values less than $\mathcal{P}$ if listeners' look at their parents when talking and when their mothers are College graduates. Factors such as whether the listener is friendly with strangers, and both the storyteller and listener use words to describe feels also led to significant differences in the distri-

[††]Trained annotators only agreed moderately while annotating these features. Refer to [35] for more details.

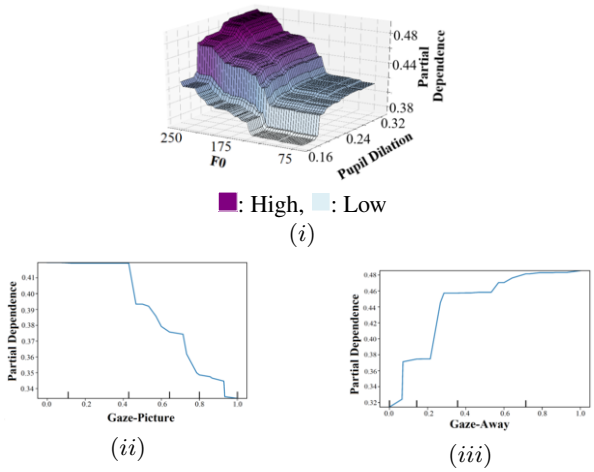


Figure 3. Partial Dependence Plots: (i) Combined effect of speaker's F0 and Pupil Dilation on backchanneling signals (ii), (iii) Impact of the proportion of time when the listener is gazing towards the picture (Gaze-Picture) and away (Gaze-Away) on listener's disengagement.

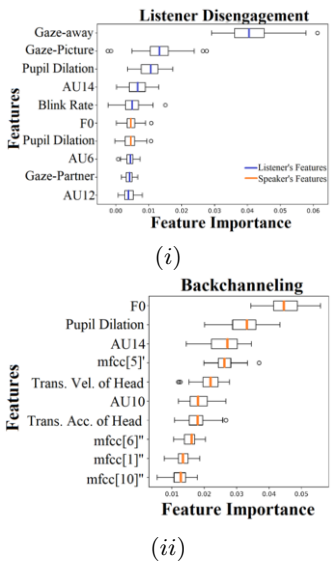


Figure 4. Permutation Feature Importance (MDA): Top 10 important features for (i) LDP and (ii) BEP.

bution of BC proportion. Overall, the socio-emotional development of the storyteller and listener in terms of their ASQ score also led to significant differences in the distribution of BC proportion. Furthermore, the fact that the gender of the storyteller influences the extent of BC probably implies that listeners respond differently to male and female storytellers. We also note that whether the storyteller and listener were of the same sex, or have siblings, had insignificant effect on the distribution of BC proportion.

Having established that some socio-demographic and developmental factors indeed influence backchanneling, we proceeded to investigate whether these features are also predictive of the extent of backchanneling. To this end, we used the socio-demographic and developmental features of the storyteller and listener and a random forest classifier to predict whether an episode will have high or low backchanneling. We labelled episodes with a BC proportion greater than median BC proportion ($\tau = 0.23$) as having high backchanneling. The detailed classification results averaged over 5-folds of

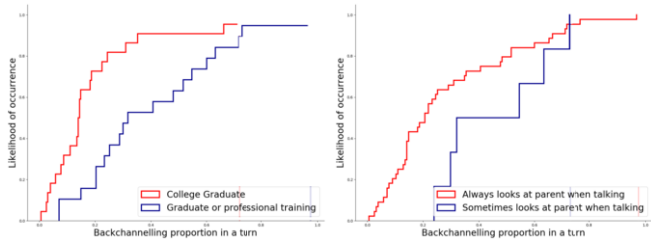


Figure 5. (i) Empirical CDFs corresponding to the listeners' mothers' education and (ii) whether they look at their parents when talking to them.

cross validation are summarized in Table 5 (i). It must be noted that our model was able to perform well (AUC $= 0.84$, F-1 ≈ 0.67) in spite of having access to only limited socio-demographic and developmental features. The most important features for the Random Forest are shown in Table 5 (ii). Highest education of listener's mother is the most important feature in predicting the extent of backchannel responses. This makes sense because mothers generally have a profound influence on a child's overall development including language and communication skills. Our results also reveal that backchannel responses depend on the gender of both the participants. ASQ score and household income of the Listener are some other features that affect backchanneling.

5-Fold Average Metrics		
Metric	**Low BC**	**High BC**
P	0.66	0.73
R	0.75	0.72
F1	0.66	0.69
AUC	0.84	

(i)

Top 5 important features
Mother's highest education (L)
Gender (S)
ASQ Score (L)
Gender (L)
House income (L)

(ii)

Table 5. (i) Predicting the extent of Backchanneling using Random Forests and socio-demographic features. (ii) Top 5 most important features for random forest model ((L): Listener, (S): Speaker).

Features	**2-sample KS Test Statistic**	
	Storyteller	**Listener**
Gender	**0.404** *	0.202
Storyteller & Listener have same gender	0.133	0.133
Mothers' Highest education	0.2129	**0.569** *
ASQ Score (Thresh = 20)	–	**0.320** *
ASQ Score (Thresh = 25)	0.320	–
ASQ Score (Thresh = 30)	**0.274** *	**0.283** *
Child uses words to describe feelings	**0.341** *	**0.318** *
Child friendly with strangers	0.146	**0.283** *
Child talks or plays with adults (s)he knows well	0.218	0.198
Child looks at parent when talking	0.273	**0.591** *
Has siblings	0.238	0.226
Total household income	0.180	**0.268** *

Table 6. Kolmogorov-Smirnov statistics. Values marked with a * indicate significant differences at 5% significance levels. The distribution of backchanneling proportion differs significantly in terms of the storyteller's gender, ASQ score, and the listener's mother's highest education *etc.*

5 Discussion and Conclusion

In this work, we focused on two roles of peer learning companions, as active storytellers and listeners, by developing models to predict listener disengagement and the extent of backchanneling. We used state-of-the-art time series classification techniques like ResNet and Random Forests to establish promising results for both the tasks. We also analyzed our results using MDI, MDA and PDPs to examine how different features impact the attentive states and backchanneling responses. Specifically, we found that for LDP the proportion of time when the listener was gazing away and towards the picture played an

important role besides other visual features such as the blink rate of his eyes. While modeling BEP, we observed that in addition to the speakers' pitch, their pupil dilation also had a positive correlation with their ability to elicit backchannel responses from the listener. Furthermore, using statistical tests and Random Forests, we found that the listeners' mothers' education and the gender of both the listener and speaker strongly influence the extent of backchanneling.

Although we aim to eventually develop engaging peer learning companions, we admit that our experiments did not include a social robot or interactive tablets. However, prior research has shown that children consider robots as social beings [21], and therefore we believe that our findings from studying peer-to-peer interactions will also apply to child-robot interactions. We also believe that further experimental validation is crucial to corroborate the effectiveness of our models.

References

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer, 'Permutation importance: a corrected feature importance measure', *Bioinformatics*, **26**(10), 1340–1347, (2010).

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, 'Openface: an open source facial behavior analysis toolkit', in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10. IEEE, (2016).

[3] Shalini Bhatia, Roland Goecke, Zakia Hammal, and Jeffrey F Cohn, 'Automated measurement of head movement synchrony during dyadic depression severity interviews', in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8. IEEE, (2019).

[4] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda, 'Social robotics', in *Springer handbook of robotics*, 1935–1972, Springer, (2016).

[5] Leo Breiman, 'Random forests', *Machine learning*, **45**(1), 5–32, (2001).

[6] Leo Breiman, 'Manual on setting up, using, and understanding random forests v3. 1', *Statistics Department University of California Berkeley, CA, USA*, **1**, 58, (2002).

[7] Hennie Brugman and Albert Russel, 'Annotating multi-media/multi-modal resources with ELAN', in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, (May 2004). European Language Resources Association (ELRA).

[8] Herbert H Clark, Susan E Brennan, et al., 'Grounding in communication', *Perspectives on socially shared cognition*, **13**(1991), 127–149, (1991).

[9] Stephanie M Curenton, 'Narratives as learning tools to promote school readiness', (2010).

[10] Iwan de Kok and Dirk Heylen, 'The multilis corpus–dealing with individual differences in nonverbal listening behavior', in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, 362–375, Springer, (2011).

[11] Alan R Dennis and Susan T Kinney, 'Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality', *Information systems research*, **9**(3), 256–274, (1998).

[12] Allen T Dittmann, 'Developmental factors in conversational behavior', *Journal of Communication*, **22**(4), 404–423, (1972).

[13] Sidney D'Mello, Rick Dale, and Art Graesser, 'Disequilibrium in the mind, disharmony in the body', *Cognition & emotion*, **26**(2), 362–374, (2012).

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller, 'Opensmile: The munich versatile and fast open-source audio feature extractor', in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, New York, NY, USA, (2010). ACM.

[15] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller, 'Deep learning for time series classification: a review', *Data Mining and Knowledge Discovery*, **33**(4), 917–963, (2019).

[16] Jerome H Friedman, 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, 1189–1232, (2001).

[17] Mononito Goswami, Lujie Chen, and Artur Dubrawski, 'Discriminating cognitive disequilibrium and flow in problem solving: A semi-supervised approach using involuntary dynamic behavioral signals', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. AAAI, (2020).

[18] Zeeshan Ali Haq and Ziaul Hasan, 'Eye-blink rate detection for fatigue determination', in *2016 1st India International Conference on Information Processing (IICIP)*, pp. 1–5. IEEE, (2016).

[19] Lucille J Hess and Judith R Johnston, 'Acquisition of back channel listener responses to adequate messages', *Discourse Processes*, **11**(3), 319–335, (1988).

[20] Dongmei Jiang, Yulu Cui, Xiaojing Zhang, Ping Fan, Isabel Ganzalez, and Hichem Sahli, 'Audio visual emotion recognition based on triple-stream dynamic bayesian network models', in *Affective Computing and Intelligent Interaction*, eds., Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, pp. 609–618, Berlin, Heidelberg, (2011). Springer Berlin Heidelberg.

[21] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Nathan G Freier, Rachel L Severson, Brian T Gill, Jolina H Ruckert, and Solace Shen, '"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, (2012).

[22] Mariska E Kret, 'The role of pupil size in communication. is there room for learning?', *Cognition and Emotion*, **32**(5), 1139–1145, (2018).

[23] Jin Joo Lee, Cynthia Breazeal, and David DeSteno, 'Role of speaker cues in attention inference', *Frontiers in Robotics and AI*, **4**, 47, (2017).

[24] Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp, and Petra Wagner, 'The alico corpus: analysing the active listener', *Language resources and evaluation*, **50**(2), 411–442, (2016).

[25] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch, 'A probabilistic multimodal approach for predicting listener backchannels', *Autonomous Agents and Multi-Agent Systems*, **20**(1), 70–84, (2010).

[26] Javier Movellan, Micah Eckhardt, Marjo Virnes, and Angelica Rodriguez, 'Sociable robot improves toddler vocabulary skills', in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pp. 307–308. ACM, (2009).

[27] Guy P Nason, 'Stationary and non-stationary time series', *Statistics in Volcanology. Special Publications of IAVCEI*, **1**, 000–000, (2006).

[28] Hiroaki Noguchi and Yasuharu Den, 'Prosody-based detection of the context of backchannel responses', in *Fifth International Conference on Spoken Language Processing*, (1998).

[29] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal, 'Telling stories to robots: The effect of backchanneling on a child's storytelling', in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 100–108. IEEE, (2017).

[30] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal, 'Telling stories to robots: The effect of backchanneling on a child's storytelling', in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 100–108. IEEE, (2017).

[31] Carole Peterson, Beulah Jesso, and Allyssa McCabe, 'Encouraging narratives in preschoolers: An intervention study', *Journal of child language*, **26**(1), 49–67, (1999).

[32] Carole Peterson, Beulah Jesso, and Allyssa McCabe, 'Encouraging narratives in preschoolers: An intervention study', *Journal of child language*, **26**(1), 49–67, (1999).

[33] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen, 'Backchannel strategies for artificial listeners', in *International Conference on Intelligent Virtual Agents*, pp. 146–158. Springer, (2010).

[34] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro, 'Automatic analysis of facial affect: A survey of registration, representation, and recognition', *IEEE transactions on pattern analysis and machine intelligence*, **37**(6), 1113–1133, (2014).

[35] Nikhita Singh, Jin Joo Lee, Ishaan Grover, and Cynthia Breazeal, 'P2pstory: Dataset of children as storytellers and listeners in peer-to-peer interactions', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 434:1–434:11, New York, NY, USA, (2018). ACM.

[36] Tereza Soukupova and Jan Cech, 'Eye blink detection using facial landmarks', in *21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia*, (2016).

[37] Keith Topping and Stewart Ehly, *Peer-assisted learning*, Routledge, 1998.

[38] Victor H Yngve, 'On getting a word in edgewise', in *Chicago Linguistics Society, 6th Meeting, 1970*, pp. 567–578, (1970).