tant aspect of the user's eye gaze is that they are looking at ERICA while she is speaking.

Although eye-gaze models using computer vision techniques have been developed in previous work (**?**; **?**), we opted to use a simpler geometry-based method with the Kinect sensor. We define the 3-d world co-ordinates of ERICA and the Kinect sensor's position, and can receive the vector of the head orientation provided by the Kinect sensor. We first transform this orientation into world co-ordinates so that we have a vector whose origin and direction are the user's head and head orientation. We then use collision detection to check for intersections between the vector and a 30cm sphere around ERICA's head. This is to accommodate for the measurement of head orientation rather than actual eye gaze. If there is an intersection, we label it as looking at ERICA.

From our experiment we found that the highest inter-annotator agreement of engagement (Spearman's correlation coefficient of 0.375) was when the subject gazed at ERICA continuously for 10 seconds during her speaking turn. Therefore, we use this rule as a basis for the eye gaze input. Turns of less than 10 seconds long were classified as negative according to the eye gaze model. We manually annotated a ground truth of eye gaze through visual observation then labelled the turn according to the 10 second rule. We then generated labels according the output of our eye gaze model. The number of turns which were labeled as positive (the user gazed at ERICA for at least 10 seconds continuously) was 17.1% of the total. The model was tested using 20 sessions of data and results are shown in Table 6.

Model	Prec.	Rec.	F1	Acc.
Baseline	0.171	1.000	0.292	0.716
Gaze model	0.504	0.580	0.539	0.847

Table 6: Performance of eye gaze detection model.

We find that the model works reasonably well for classifying continuous gaze behavior. However we could only estimate the positions of ERICA's head and the Kinect sensor in our corpus by observing the video. In the live system we calculate these values exactly, so we expect that the performance of the model will be improved from this result.

Engagement Recognition

We have described four different social signals for and their recognition models. The performance of the models is varied, but our goal is not to produce state-of-the-art individual models, but to assess whether they can be used in conjunction with our engagement recognition model. This is a hierarchical Bayesian binary classifier, predicting if a listener is engaged or not during the system's speaking turn. The graphical model is shown in Figure 3.

For our evaluation we selected 20 sessions from our corpus which were a different subset than those used to train the individual social signal models. We recruited 12 third-party observers to watch video of these sessions and annotate time-points where they perceived the engagement of the subject was changed to high. We defined engagement for the annotators as "How much the subject is interested in and willing to continue the current dialogue with ERICA". The annotator variable is represented by i in our graphical model.

The input to the model, x_t is a state which is defined as the observed combination of social signals during a turn. Each social signal can be classified as a binary value (either detected or not detected), giving 16 possible combinations. When training the model we know this combination, but in a live system we marginalize over each behavior combination using its prior probability.

We also define a variable k, which represents a specific character type of an annotator. For example, laughter may be influential for one character type but not so important for another. Therefore, annotators with different characters would perceive laughter differently in terms of perceived engagement. We created a distribution of characters for each annotator and found that they could be grouped by similar character. From our experiments we found that including a character variable improved the model's performance and that the best model had three different character types ($K = 3$).

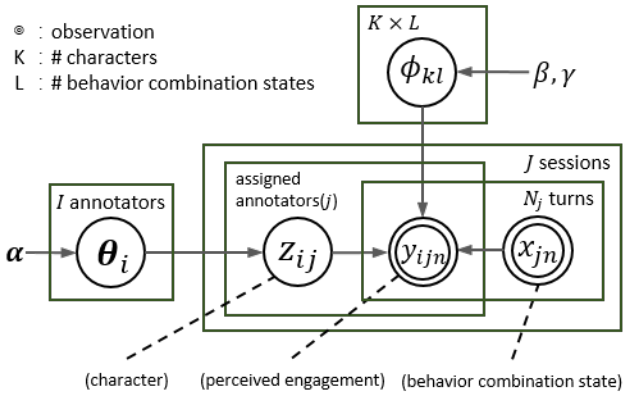

Figure 3: Graphical model of engagement recognizer

For brevity, we omit details of most of the calculations used in our model. The posterior probability of perceived engagement is

$$p(y_{it}|x_t, i, \tilde{\Theta}, \tilde{\Phi}) = \sum_{k=1}^{K} \tilde{\theta}_{ik} \tilde{\phi}_{kx_t} \quad (1)$$

where x_t is the observed behavior combination for a system turn, i is the index of the annotator, and Θ and Φ are the learned parameters of the latent character model. $\tilde{\theta}_{ik}$ is the probability that an annotator i has a character k, and $\tilde{\phi}_{kx_t}$ is the probability that the behavior combination x_t is perceived as engaged by k.

We compared the performance of our engagement model under two conditions - using manually annotated data as inputs and using the results of our social signal detection models as inputs. Engagement is classified per the robot's speaking turn. We also analyze the model according to whether it uses contextual information. This means the result of engagement in the system's previous turn is used as a fea-

ture for classifying engagement in her current turn. We use the area under the precision-recall curve (AUC) as a performance measure. The results are shown in Table 7.

Labeling method	No context	Context
Manual annotation	0.650	0.669
Detection system	0.615	0.620

Table 7: AUC scores for the engagement model using manual annotation and our social signal detection models.

From our results we see a drop in performance of the engagement model when using social signal detection compared to manual annotation, but is not drastic. We also see that adding contextual information provides an improvement in performance. We can also show that even though the individual models do not all have high recognition ability, their combination is adequate for engagement recognition.

Discussion

We find that the individual social signal models have varying levels of performance. Laughter detection is quite poor, while backchannel detection is considerably better. To improve the performance of our recognition systems, we could add other modalities as in other works (**?**). In particular, the co-gesture of verbal backchannels and nodding could help to improve both systems. Similarly, visual recognition of laughing would improve the results of laughter detection. We are trying to improve the performance of the individual models, including using spectral features for better laughter and backchannel detection.

Although we have evaluated the model using our corpus, we expect that laughter and backchannel detection performance in the live system will be slightly degraded because these signals are mixed in with ERICA's speech during her turn. We limit this by using the microphone array to ignore ERICA's voice, but cannot guarantee that user speech will be clean. Other previous works tend to focus on non-verbal social signals, but from our third party annotation experiment, verbal signals are necessary.

Our data was collected in a one-to-one conversational setting, but the models are not restricted to this specific environment. Nodding, laughter and backchannel detection are independent of both environment and user. The eye gaze model needs to be calibrated to accommodate the position of Kinect and ERICA. We have successfully implemented all our models in a separate environment, with different placements of Kinect and ERICA. We propose that the models can function in a varied number of conversational settings, including multi-party dialogue.

Our next step is to use the results of engagement recognition to modify the dialogue policy of the system. We consider that the engagement of the user has an influence on turn-taking behavior or changing the topic of conversation. However, we can also consider that the flow of dialogue may be completely modified. One scenario we are considering is ERICA giving a technical explanation. By recognizing if the user is engaged, the robot may use simpler terminology to make her talk more understandable. We intend to formulate such scenarios where conversational engagement recognition is necessary and then conduct user experiments to confirm the effectiveness of our system.

Conclusion

This paper described models for detecting nodding, laughter, verbal backchannels and eye gaze, which will be used by an engagement recognizer during conversation with a robot. The robot we use in this work is the android ERICA. We selected these social signals based on a previous experiment where third party observers annotated changes in engagement based on behaviors. The inputs are a Kinect sensor and a microphone array. Although the performance of our models are varied, their combination is effective. We observe a slight degradation in performance for our engagement recognition model when using the outputs of social signal detection compared to annotated values. We have integrated these models into ERICA's system architecture and intend to make her an engagement-aware conversational robot.

Acknowledgements

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401), Japan.

Nesciunt quae illum dolore animi delectus veritatis quisquam, sunt fugit accusantium atque natus quod itaque at eveniet, quo esse cum recusandae magnam autem, ipsam laboriosam ullam eligendi quo architecto expedita iste.Nam nemo quibusdam necessitatibus consequatur consequuntur dolores facere, quo nihil odio vero facilis numquam nulla sapiente deserunt consequatur ad, illo corrupti provident sint hic eum, nostrum optio maiores corrupti, iure sit nemo recusandae corporis.Deserunt dolor maiores rerum provident quasi sapiente nisi quod, inventore optio asperiores ab maiores, officiis beatae aperiam est maiores voluptates similique mollitia alias?Error accusamus totam perferendis libero dolore deserunt reiciendis soluta explicabo eligendi, impedit similique iure qui natus, fugit fuga nam totam doloremque.Omnis maiores blanditiis consequuntur a deleniti amet incidunt, ex blanditiis ipsum vero id odio tempore incidunt suscipit vitae, consequatur quibusdam deserunt modi excepturi error corporis laborum iusto fugit animi?Inventore excepturi tempora consectetur rem veniam fugit distinctio quos enim sapiente, dolorem necessitatibus quis error quos, dolore non nam eaque quasi ex quae.Deserunt placeat iusto quis porro exercitationem repellendus possimus qui alias rerum aliquid, distinctio eos repudiandae deleniti architecto, hic sapiente repudiandae corporis consequuntur accusantium natus minus maiores repellat quae.Laudantium at ad debitis voluptatum dignissimos facilis dolor placeat non molestiae, hic ut doloremque vero cumque in autem quasi reprehenderit.Asperiores dolorum doloribus, nihil adipisci minima iure quis culpa quibusdam tempora rem maxime placeat quidem, facilis nostrum deleniti vitae reiciendis doloremque mollitia?Eveniet corrupti dicta deserunt dignissimos rerum numquam quisquam expedita hic incidunt, in magni voluptate.Ipsa voluptates debitis neque sint velit numquam alias ratione in temporibus sapiente, sapiente deleniti pariatur enim suscipit itaque distinctio, maxime incidunt perferendis aspernatur voluptatum voluptatibus error