Backchannel Generation Model for a Third Party Listener Agent

Divesh Lala
Graduate School of Informatics, Kyoto University
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Graduate School of Informatics, Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Graduate School of Informatics, Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

Kei Sawada
rinna Co., Ltd.
Tokyo, Japan
keisawada@rinna.co.jp

ABSTRACT

In this work we propose a listening agent which can be used in a conversation between two humans. We firstly conduct a corpus analysis to identify three different categories of backchannel which the agent can use - responsive interjections, expressive interjections and shared laughs. From this data we train and evaluate a continuous backchannel generation model consisting of separate timing and form prediction models. We then conduct a subjective experiment to compare our model to random, dyadic, and ground truth models. We find that our model outperforms a random baseline and is comparable to the dyadic model despite the low evaluation of expressive interjections. We suggest that the perception of expressive interjections contribute significantly to the perception of the agent's empathy and understanding of the conversation. The results also show the need for a more robust model to generate expressive interjections, perhaps aided by the use of linguistic features.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

dialogue systems, multiparty dialogue, listening agent

ACM Reference Format:
Divesh Lala, Koji Inoue, Tatsuya Kawahara, and Kei Sawada. 2022. Backchannel Generation Model for a Third Party Listener Agent. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3527188.3561926

HAI '22, December 5–8, 2022, Christchurch, New Zealand

ACM ISBN 978-1-4503-9323-2/22/12…$15.00
https://doi.org/10.1145/3527188.3561926

1 INTRODUCTION

Recent conversational agents have become more sophisticated, largely through advances in deep learning models for natural language processing [2, 21, 28]. While these advances allow the agent to become a better speaker, contributing to the conversation as a listener is arguably just as important. One way to do this is to give the agent the ability to produce appropriate backchannels as listening behavior. In this work we focus on verbal backchannels which include (in English) *hmm, yeah* and *uh-huh* - short lexical tokens which do not interrupt the speaker but react to the conversation.

While backchannels have been mostly implemented in agents which engage in dyadic conversation [6, 11, 24], the same phenomena should be addressed in group conversations, particularly with the use of online platforms to facilitate multiparty chatting. A conversation with multiple speakers may be more difficult to parse for an agent due to speech recognition and natural language processing issues involved with overlapping talk. However, it may still be possible for the agent to contribute as a listener, providing appropriate feedback to the conversation without taking any turns. The agent should participate in a conversation between two or more humans by providing attentive backchannel feedback.

In this work we implement such an agent which has a role as a listener during a talk between two human participants. The agent uses only the prosodic audio information of the speakers (i.e. no speech recognition) and so can function in real-time with little latency. The agent generates backchannels throughout the conversation according to a trained model. Although our agent is implemented in the context of a radio show and is not embodied, we propose it can also be integrated in embodied agents and robots for situated group conversation. In this work the target language is Japanese. Figure 1 shows the overall concept.

This paper contains several research contributions. Firstly, in Section 3 we analyze a corpus of radio show sessions with a third party listener to identify appropriate backchannel categories which can be used by the agent. We then train and evaluate a continuous backchannel generation model in Section 4 which provides us with an objective performance metric.

In addition to the objective evaluation, we also conduct a listening experiment in Section 5 where the proposed agent is compared to others in terms of subjects' perceptions of its empathy and understanding, both important features for agents [12, 17]. Subjects also directly evaluate each individually generated backchannel, giving an insight into how the rating of a single backchannel contributes

Table 1: Categorizations of third party backchannels in the initial corpus based on Den et al. [5]

Category	*Function*	*Examples (including English)*	*Proportion(%)*
Responsive interjections	Show acceptance of another utterance	un, unun, hai (mm, mmhmm, yeah)	22.2
Expressive interjections	Expresses admiration, surprise disappointment etc. elicited by another utterance	u-n, o-, a- (huh, wow, aah)	26.3
Repetitions	Repeating a portion of what another participant has said to express understanding or agreement	A: otagai issho dattane -> B: **issho** dayo (A: do it together -> B: **together**)	11.2
Shared laughs	Laughing with another participant	-	11.8
Other	Backchannels which don't fit into one of the above categories	honto desu ka (really?)	28.4

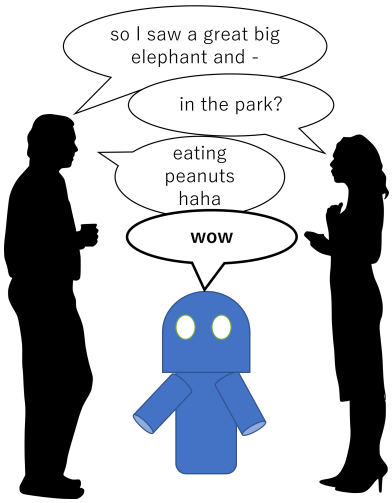

Figure 1: Concept of multiparty backchannel system

to a user's overall perception of an agent. The paper then concludes with a discussion and future directions.

2 RELATED WORK

Backchannel prediction, both verbal and non-verbal, has received much attention in order to make listener behavior of an agent more humanlike [10, 16, 23, 25, 26]. Integration of these systems into intelligent agents has also been implemented in several works, with special care needed to regulate their frequency, timing and type [4, 20]. Well-known examples of these include Sensitive Artificial Listeners [24] and SimSensei [6].

Features used to make predictions are often prosodic as low latency is preferable for real-time implementation, however these can differ according to language. For example, Ward and Tsukahara [26] found low pitch to be a robust predictor for Japanese backchannels rather than simply waiting for silence, while Truong et al. [25] found that the duration of a pause is important for Dutch conversation. Advances in speech recognition has meant that word embeddings have also been explored [19, 22]. Backchannel prediction is not only about timing, but the morphological form of the backchannel [20]. For Japanese, this type of prediction has been attempted in several works [1, 3, 11, 13, 27], however results are still modest (F-score around 0.5), showing the difficulty of this task.

Furthermore, while much research focuses on prediction in a dyadic context, our work will predict backchannels in a multiparty conversation, where the agent acts as a third party listener rather than an active speaker. Although multiparty listener behavior has been analyzed in terms of gaze and nodding [9, 14], we could only find one work which implemented verbal backchanneling in a multiparty context [18]. In this work, a rule-based backchannel model using gaze information was implemented, although was not independently evaluated.

In our work, we train a model for Japanese verbal backchannelling in a multiparty context, which has not yet being addressed. Although it is possible to simply use existing dyadic models, it is also known that dialogue behaviors change depending on group size [7] and similarly prosodic rules in dyadic models may differ in multiparty chat. In this work we investigate this by directly comparing our model to one trained on dyadic conversation. Furthermore, we argue that user perception of backchannels is required for model evaluation, since the timing and form of backchannels is not an objective truth. Many works provide only objective measures as evaluation. We also provide a rigorous analysis of subjective data in which not only the overall conversation but the individual backchannels themselves are assessed. As far as we know this type of analysis has not been conducted in any previous work.

3 DATA COLLECTION AND ANALYSIS

We considered data collection in the context of the agent implementation. Although it would have been possible to use three-way conversations, our target agent does not play the role of a speaker, and so using data where a third party acts as both speaker and a listener would not align with this. However we do not have a public corpus with a third party only speaking in backchannels. To deal with this issue we made the decision to augment conversation (a radio show), with a third party listening to two speakers. Although this is a limitation, we argue that data that we gather would better reflect the intended role of the listening agent, which uses backchannels but does not directly intervene in the conversation.

In this work we use data from a public radio show with two young female radio hosts. The show consists of light-hearted casual chatting between the hosts on various topics. A third party listener,

Table 2: Distribution of categorical forms and tokens in the second corpus

Category	*Frequency*	*Percentage (%)*
Responsive tokens	**1147**	**45.9**
un	488	19.5
hai	222	8.9
un un	151	6.0
un un un	118	4.7
ha-i	72	2.9
other	96	3.8
Expressive tokens	**908**	**36.3**
u-n	234	9.4
ne-	168	6.7
a-	106	4.2
he-	85	3.4
o-	80	3.2
e-	37	1.5
other	198	7.9
Shared laughs	**444**	**17.8**
Total	**2499**	**100.0**

also female, was recorded responding while listening to the pre-recorded sessions. This person obviously could not interact with the radio hosts so they were restricted to reacting to what was being said rather than adding information to the conversation. A total of 17 data sessions were analyzed, with an average length of 258 seconds each. Our goal of this initial corpus analysis is to only identify the major types of backchannels used by the third party listener, which will then be implemented for our agent. In Section 4 we will use different sessions for the actual model training.

There were a total of 1185 transcribed utterances, which we annotated and categorized. We used previous literature on Japanese backchannels [1, 5, 26] to identify four major categories - responsive interjections (responsives), expressive interjections (expressives), repetitions and shared laughter. Other types of utterances for which the category was not clear were labeled as "Other". Although it is possible to break down the backchannels into a larger number of categories, this would make model training more difficult. Table 1 shows the categories, their function, and some examples.

After consideration of the agent model, we decided to omit the repetition category, since we do not have reliable speech recognition. Therefore the three target categories of backchannel used in our listener agent are responsives, expressives and shared laughter, comprising of approximately 60% of all backchannels used by the third party listener. Responsives act as an indication that the listener is attending to the conversation while expressives are more emotional, indicating surprise or interest in what has been said. Their use is more dependent on the context of the conversation.

4 MODEL IMPLEMENTATION

The results from our data collection motivate us to train a model to classify three main categories of backchannels for our listener agent: responsives, expressives and laughs.

4.1 Training data

Although we could use the same radio sessions in the previous section to train the model, 40% of the utterances used in the data were not covered by the three categories. We decided to create a second corpus of 24 sessions where the third party listener uses *only* the three target categories of utterances. The sessions in this corpus are all different from the previous one. Another third party listener undertook the same task as described previously, except they were instructed to only use backchannels which fell under one of the three target categories. An example dialogue (English translation) was as follows:

Host A it's a banana wrapped in something like sponge
Host B oh yeah like sponge
Listener wow (expressive)
Listener i see (responsive)
Host B you mean actual sponge
Host A ah right *laugh*
Listener *laugh* (shared laugh)

Table 2 shows the distribution of the target categories and individual backchannels in this second corpus. Interestingly, this corpus contained a greater number of responsives than expressives, the opposite of the initial corpus. One explanation for this could be that when given restrictions on the utterances which can be used, the third party listener opted for responsive backchannels as they could be applied to a variety of situations without being inappropriate. The distribution of these backchannels also differs from other analyses of dyadic conversations in Japanese [15, 26]. We note the high frequency of "other" expressives. Most of these backchannels appear less than three times in the entire corpus.

We also assume that backchannels within the responsive category are more interchangeable than those in the expressive category. For example, replacing *un* with *un un* in most cases changes the semantics slightly but does not feel inappropriate. On the other hand, for expressives such an exchange might result in unnatural listening behavior. We can imagine in English that "huh" and "wow" are not readily interchangeable because they have different meanings. This has implications for our subjective experiment.

4.2 Model architecture

One major consideration when designing our agent model was how to predict both the timing and the categorical form of the backchannel. We had two options. The first option is to predict both timing and form together. This uses four classes (the three target categories plus a silence class) with the model making a decision based on one continuous prediction. The second option is to use a dual model approach as shown in Figure 2 - a binary model which first predicts if a backchannel is used and, if positive, a 3-class model which then predicts which backchannel form should be generated. We tested both these architectures and found that the second approach was more favorable as the first approach suffered from class imbalance which negatively affected its performance.

4.3 Labeling and features

Before the model training we carefully considered the type of features to be used, in particular whether to include linguistic features. Our intention is for the model to be usable in a real world context

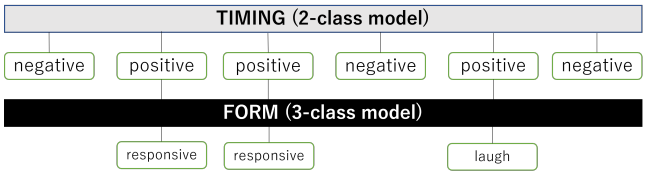

Figure 2: Dual model architecture used for prediction of backchannel timing and form

so we cannot assume speech recognition for both parties is always available. In fact, the training data itself is only a single combined channel with both speakers. We should also not assume that speech recognition will be accurate, particularly since the radio show contains much informal language with disfluencies and interruptions. Furthermore, using linguistic features assumes that backchannels are made in reaction to an utterance. However, from our data we observed that backchannels often occur within speech and before an utterance has been completed. An obvious example is shared laughs, where it makes little sense to wait until the other party has concluded their laugh utterance before responding.

Given these limitations and the need for the performance of the final model to fairly reflect the training environment, we decided to omit linguistic features in this work, although they would arguably improve the model if used. On the other hand, our model is lightweight, only requiring one channel of streaming audio data and voice activity of each speaker to be functional in real-time.

Figure 3 summarizes the labeling and feature windows. If we consider positive labels at only the time point at which the backchannel is uttered, there will be a massive class imbalance. Therefore we used a labeling window where all time points within 500ms before and after the beginning of the backchannel are positive. For all time points within this labeling window, the training samples for the timing model will be positive and used for the form model, labeled as the corresponding form category. Occasionally these timepoints may overlap with multiple labeling windows if different backchannel forms were uttered quickly one after the other.

For a particular timepoint, we extract a feature set of prosodic information using a pitch extractor [8]. These features are taken from feature windows with lengths of 100ms, 200ms, 500ms and 1000ms directly before the timepoint. Audio data within them is sampled every 10ms. For each window we calculate these features:

- median pitch and power
- percentage of the window which is voiced (i.e. contains a pitch)
- percentage of silence from both radio hosts

This gives us 5 features over 4 feature windows, or 20 features per training sample. In our data collection environment, pitch and power information is a single combined audio channel of both hosts, which simplifies the model. We estimate the voice activity information of each individual host through transcripts to extract the percentage of silence, though voice activity detection can also be achieved in real-time.

Table 3: Performance of timing prediction model

Model	*Precision*	*Recall*	*F-Score*
Random baseline	0.299	0.299	0.299
Proposed model	0.393	0.601	0.475

Table 4: Performance of form prediction model. The bottom table reports macro scores compared to the baseline.

Class	*%*	*Precision*	*Recall*	*F-Score*
Responsives	0.462	0.598	0.724	0.655
Expressives	0.369	0.462	0.249	0.324
Shared laughs	0.169	0.433	0.620	0.510

Model	*Precision*	*Recall*	*F-Score*
Random baseline	0.333	0.333	0.333
Proposed model	0.498	0.531	0.496

4.4 Model performance

We used the above samples to train logistic regression models as the machine learning classifiers. Deep learning techniques such as RNNs were also tried, but were found to not improve performance. We assigned each of the 24 sessions as a fold and trained the models using leave-one-out cross-validation. Predictions for the timing model were made every 50ms for a total of 165,795 samples. Predictions for the form model were only considered if the ground truth label was classified as being a backchannel, with a total of 48,481 samples used. We compared both timing and form models to a random baseline which uses the class distribution to make the prediction. Results are shown in Tables 3 and 4.

The proposed timing model outperforms the baseline, although the F-score is still modest. However, this may still be reasonable considering that the assessment is made against continuous time points. For the form model, responsives and shared laughs are classified reasonably well, however expressives are classified at no better than the baseline rate, specifically the low recall score. Despite this, the whole model does outperform the baseline.

The major caveat when assessing the performance of this model is that although we use the"ground truth" for the correct labels, it is not completely objective. A different person will almost certainly speak with different timings and backchannel forms. Therefore a subjective experiment is needed to assess the model's real performance according to human listeners.

4.5 Backchannel generation

The model as it is does not have any decision-making processes on when the agent should actually speak and only makes predictions every 50ms. It is inappropriate to have the agent speak at the first positive prediction as this could be an outlying false positive. Therefore we designed two decision-making heuristics to regulate the agent backchannels. The first is the number of consecutive positive predictions which must be received before the agent generates the backchannel. We set this to 5 predictions (250ms), meaning that the agent will only use an utterance if the timing model receives 5

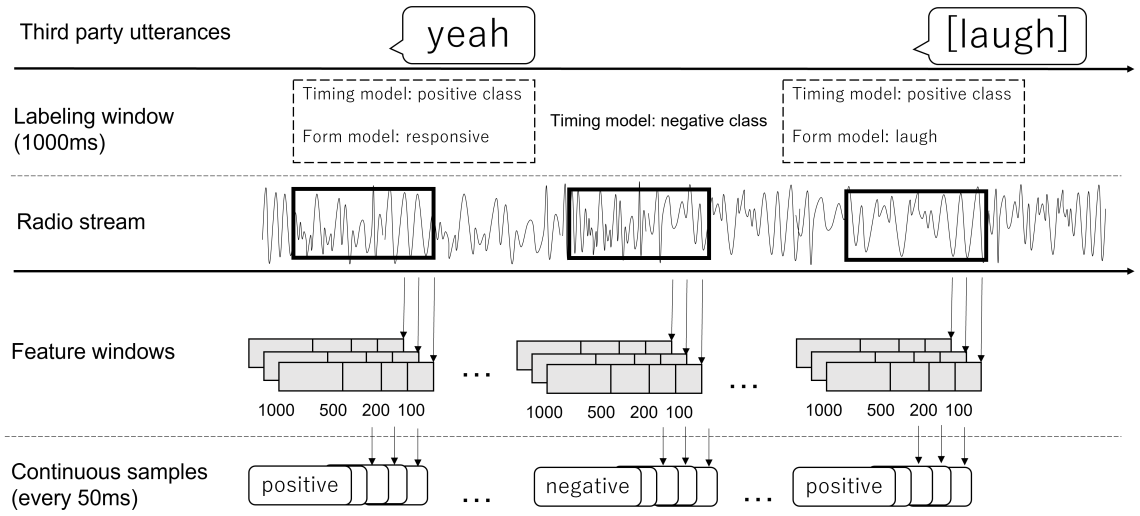

Figure 3: Overview of labeling and feature extraction.

consecutive positive predictions and the form predictions are all from the same category. The other heuristic is the amount of time to wait between backchannels, which we set at 1000ms so the agent does not use them too frequently.

We must also make a decision about which specific utterances within each form category to use. In this work a text-to-speech (TTS) system generates these utterances. Since no other information is available, we simply choose the utterance within the form category at random based on the corpus distribution presented in Table 2. It should also be noted that each utterance (and laughs) has several TTS variations. For example the utterance *un* has variations with slightly different pitches. These are selected at random to ensure the agent's utterances do not become monotonous.

5 SUBJECTIVE EVALUATION

We conducted an experiment to assess the subjective performance of our proposed model (**Proposed**) in comparison to three other models. All use the same TTS system:

- **Random:** A model which generates backchannels with random timing according to the distribution used for the binary model and random form according to the distribution in Table 2. The beginning of an agent's utterance occurs at a minimum of 1500ms before the start of the next one.
- **Dyadic:** A backchannel model we implemented in previous research [11] which was trained on dyadic conversations and only for responsive forms. We use this as a comparison because it represents a robust but conservative type of model which is used for more common dyadic interactions. Since it only uses responsives any errors will arguably be not as critical as a model which incorrectly uses expressives.
- **Ground Truth:** A model in which the actual utterances recorded from the human third party listener are directly replaced with utterances generated by the TTS system.

5.1 Sample generation

We created audio samples for the experiment by using the models to generate agent utterances for every session and used segments as experiment samples based on the following criteria:

- The first agent backchannel does not begin within the first four seconds of the sample, to give the subject some initial context to the conversation
- The length of the sample is less than 30 seconds in length
- The number of backchannels generated within the sample is between 4 and 8
- For the proposed model, the sample must contain at least one of every form.

These criteria were designed to allow the subject to listen to the conversation and each of the backchannels in context to help them best evaluate it. Additionally, we wanted the subjects to listen to as many different samples as possible so tried to restrict the length of them. It was not possible to have subjects listen to the same segment generated with different models, since there were few segments where every condition met all the criteria. Instead, we ensured every sample was non-overlapping and only treated with a single condition. In total we generated 101 total samples - 30 proposed, 20 random, 27 dyadic and 24 ground truth.

5.2 Procedure

We designed software which allowed subjects to listen to an audio sample while rating each backchannel. The software displayed the agent backchannels on an audio timeline (so subjects were aware when the agent would speak) and subjects marked each backchannel as being appropriate, somewhat appropriate or inappropriate as a direct evaluation. They also rated the sample overall for measures of empathy and understanding with the questions *"In this recording, how much did you feel that the system showed [empathy/understanding]?"*. These ratings were done on a 5-point Likert scale.

A total of 33 subjects participated in the experiment, 14 female, all university students. The subjects listened to 32 samples in total - 8 from each condition. The selection of samples for each condition was random as was the order they were presented. Subjects could listen to the samples as many times as they wished.

6 EXPERIMENT RESULTS AND ANALYSIS

We conducted three types of analysis of the subjective ratings from the experiment. Firstly, we analyze the ratings of each individual backchannel in the samples. Then we analyze the samples' overall ratings of empathy and understanding. Finally, we link the two by analyzing how individual backchannels influence the overall rating.

6.1 Individual backchannel appropriateness

Individual backchannels were given one of three ratings - inappropriate, somewhat appropriate and appropriate. We calculated the percentage of backchannels for each of these ratings according to model type and form category. Chi-square tests of independence were also calculated for each pair of model types. These tests showed significant differences for every pair of distributions except one - expressives in the proposed and random models. Figure 4 displays these graphs.

As expected, the ground truth model was the best performing overall. However approximately 10% of the backchannels were still deemed to be inappropriate. The main contribution of this is the expressives, in which about 20% of backchannels generated with the ground truth model were inappropriate. Furthermore, only about half the backchannels were deemed appropriate, suggesting that subjects rated expressives more harshly.

The proposed model outperformed the random and dyadic models in terms of the overall percentage of appropriate backchannels. When compared to the dyadic model, it received a greater proportion of both appropriate and inappropriate responsives. For expressives it was no better than the random model, with only around 20% of this type of backchannel being deemed to be appropriate. The best performing form category was laughter, with over 80% of laughs being rated as appropriate. In general laughter backchannels were rated high by subjects, with even 50% of the random model laughter backchannels receiving an appropriate rating.

6.2 Empathy and understanding

For each sample, subjects also evaluated the overall empathy and understanding of the system on a Likert scale from 1 (lowest) to 5. Every subject rated 8 samples from each of the four conditions. We show the distribution of ratings in Figure 5.

We note the relative infrequency of "neutral" opinions (i.e. Likert scale rating of 3), showing that subjects tended to have an opinion on whether a sample showed empathy and understanding. The ground truth model performed the best while the random model performed the worst. Samples generated by the proposed model were generally rated better or equal to the dyadic model samples.

We further analyzed the results in terms of the percentage of samples where the ratings of empathy and understanding were "positive" (4 or 5) or "negative" (1 or 2), shown in Table 5.

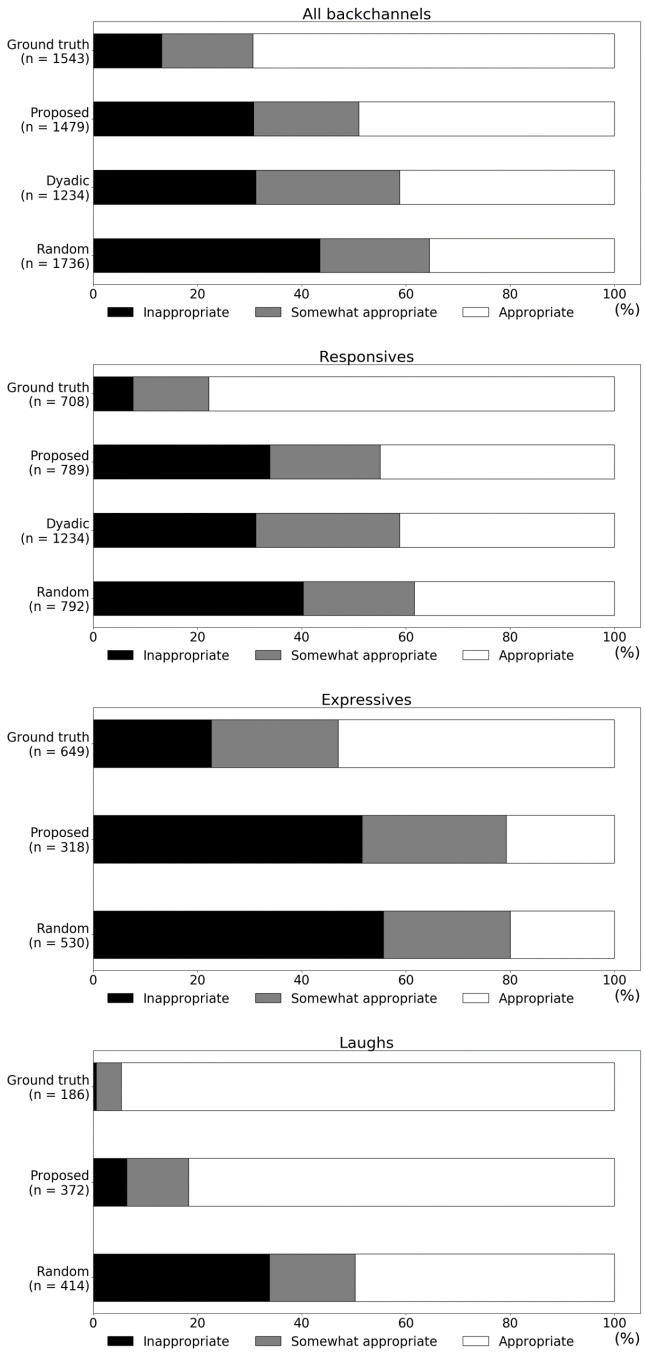

Figure 4: Percentage of backchannel ratings according to model type for each form category

The proposed model is slightly better than the dyadic model, although the difference is modest. It is also noticeable that in general subjects rated understanding lower than empathy except for the ground truth model. This perhaps indicates that creating a

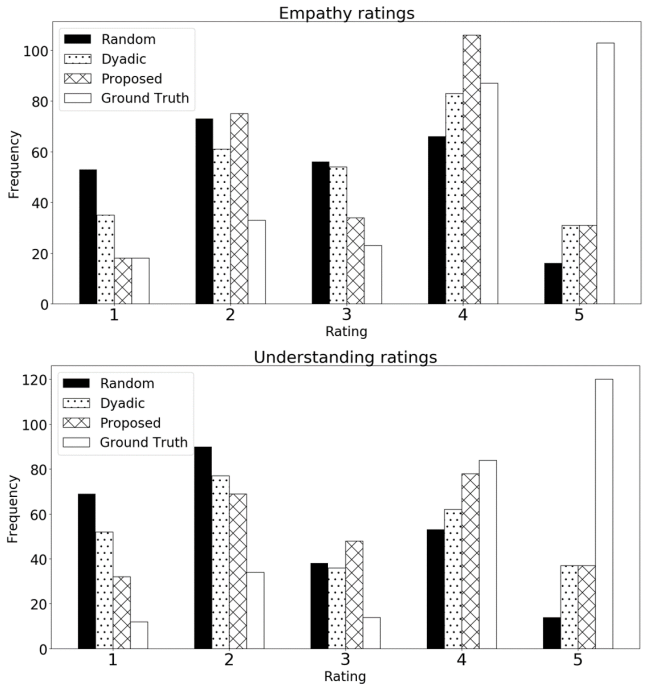

Figure 5: Frequency distribution of Likert scale ratings for empathy (top) and understanding (bottom)

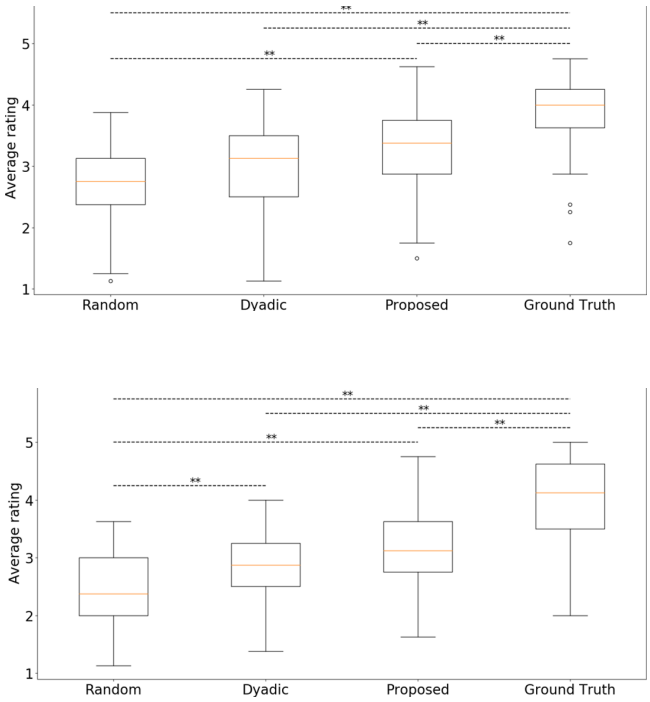

Figure 6: Boxplots of mean averages of empathy (top) and understanding (bottom) for a session. ** shows $p \leq 0.05$.

Table 5: Percentage of samples with positive and negative ratings

	Empathy		*Understanding*	
Model	Positive	Negative	Positive	Negative
Random	31.1	47.8	25.4	60.2
Dyadic	43.2	36.4	37.5	48.9
Proposed	51.9	35.2	43.6	38.3
Ground truth	72.0	19.3	77.2	17.4

backchannel model that shows understanding is more difficult than creating an empathic one.

We also performed a subject-targeted boxplot analysis. For each condition, every empathy and understanding score was averaged across the 8 samples that each subject listened to. These averages per session were compared to each other using an ANOVA test which was found to be significantly different ($p < 0.001$). We then performed Mann-Whitney post-hoc tests to compare each group, with a Bonferroni corrections used for multiple comparisons. Results are shown in Figure 6.

For both empathy and understanding, the proposed model was not significantly different than the dyadic model, although the median average rating was higher. Medians were slightly lower in understanding than empathy, except for the ground truth model.

Although our model outperformed the random baseline, it is still on par with the dyadic model. When considering the reasons for this, the obvious conclusion is that the relatively poor performance in generating expressive backchannels has a negative effect on the proposed model. It performed about the same as a random one in terms of the rating of expressives. However, it also suggests that the negative influence of expressives is counterbalanced by the better performance of responsive backchannels and laughs, which allowed it to outperform the random model overall. The next section looks a little deeper into these relationships.

6.3 Odds ratio analysis

We wanted to further understand how the ratings of individual backchannels contributed to the empathy and understanding ratings of the samples. More specifically, if a subject rates a backchannel as being appropriate (or inappropriate) how much does this increase (or decrease) the likelihood that they will also rate empathy and understanding higher?

For this purpose we trained ordinal logistic regression (OLR) models, since we have a natural ordering of Likert scale data. The input variables were the total number of inappropriate, somewhat appropriate and appropriate ratings for each sample, with the output variable being the rating of empathy or understanding. We omitted data of the dyadic model since this did not contain expressives or shared laughs. This provided us with a total of 792 samples to train the models. The overall prediction accuracy for the 5-class model was 55.7% for understanding and 52.1% for empathy.

The exponential of the coefficients of the OLR models can be used to provide an odds ratio for all categories of backchannel or each individual form category. Put simply, the model gives the odds that the empathy or understanding rating will increase for each

type of backchannel rating (inappropriate, somewhat appropriate and appropriate). If this odds ratio is close to one, it suggests that an increase in the specific rating of the backchannel has little effect on empathy or understanding. Figure 7 displays the results.

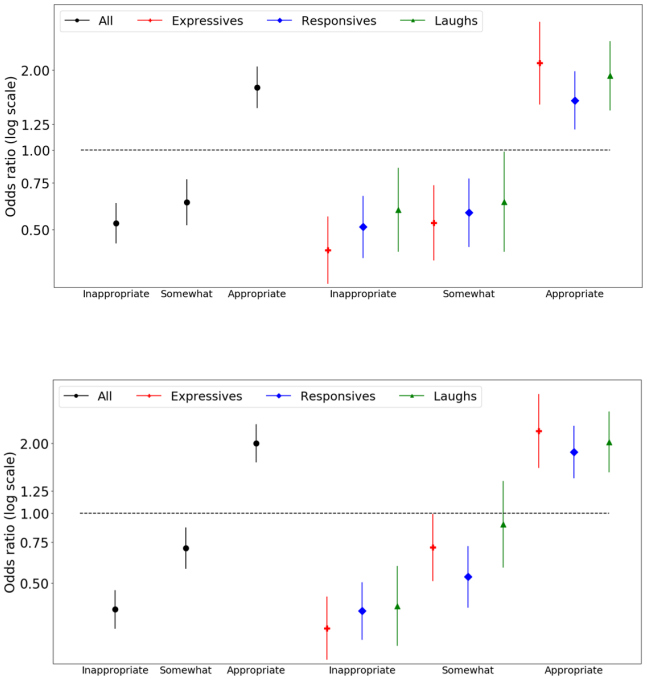

Figure 7: Odds ratio analysis for metrics of empathy (top) and understanding (bottom)

When considering all backchannels, one which is rated appropriate increases the likelihood of increasing the empathy and understanding score by around 50-100%. Conversely, a backchannel that is rated inappropriate reduces the likelihood of this by around 50-65%. For empathy there is little difference between inappropriate and somewhat appropriate backchannels, but there is a difference between these two for the empathy metric.

For backchannels in individual form categories, we see that there are no striking differences. There is a suggestion that an inappropriate expressive affects the metrics more than inappropriate responsives and laughs, but is not statistically significant. One interpretation is that using inappropriate expressives is less conservative and more noticeable to participants.

There is a suggestion that there is little effect of a somewhat appropriate laugh on understanding. One possible interpretation is that using laughs which are not completely appropriate does not show a lack of understanding, rather that subjects thought the system might simply be showing general enthusiasm.

7 DISCUSSION

In this paper we trained and implemented a third party listening agent and tested it in both objective and subjective experiments. Our proposed agent model could outperform a random baseline, was comparable to a dyadic model which only used responsive interjections, but was still far from a ground truth model. This paper has highlighted the importance and difficulty of expressive interjections as backchannels. Our model did not perform any better than random at generating these, and even in the ground truth model expressives were the worst performing.

In order to generate a more robust model for expressives, it is clear that using only a prosodic approach has limits. In this work we were interested in creating a real-time model with low latency and without speech recognition, so we restricted ourselves to using continuous features. However, including linguistic features and natural language understanding techniques would have to be the next goal for correctly classifying expressives. Our future plan is to better classify expressives using linguistic approaches, while maintaining the continuous aspect of the model. The use of low-latency incremental speech recognition is required here.

On the other hand we also found that the dyadic model which only generates responsives is still quite useful. Using only responsives in Japanese seems to have some effect on the perception of the agent. However we do not know if we have reached a limit in the model's effectiveness. If this is the case, then adding more variety to the backchannels (i.e. expressives and laughter) would improve the model, as long as we can guarantee that these backchannels could be robustly generated. For our proposed model, the improvement from increased variety and performance of laughter prediction were counteracted by weak expressive prediction. This resulted in it not significantly outperforming the dyadic model.

There are several limitations to this work. As discussed, the data we used does not actually have the third party listener interacting with the other two participants in the conversation. This could have a large effect, particularly when we consider the different dynamics of dyadic and multiparty conversations. We do not know how this model will scale up with three or more humans since we use individual voice activity information. Although we only used information from one audio channel, we might improve performance by using prosodic features from individual speakers. Furthermore, training sessions used the same female radio hosts and third party speakers.

8 CONCLUSION

This paper presented a third party listener agent which generates backchannels during a conversation between two human participants. We trained the model based on a corpus which identified three different categories of backchannel - responsives, expressives and laughs and predicts backchannel timing and form category. We conducted a subjective experiment to compare the model with three others and found our agent model could outperform a baseline model but was not significantly different than a dyadic model which only used responsive backchannels. Our analysis showed that although our model could reasonably predict laughs and responsives, its performance was degraded by relatively poor expressive prediction. We conclude that robustly predicting and generating expressive backchannels needs to be investigated further.

ACKNOWLEDGMENTS

This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011.

REFERENCES

[1] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards Immediate Backchannel Generation Using Attention-Based Early Prediction Model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7408–7412.

[2] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[3] Nicola Cathcart, Jean Carletta, and Ewan Klein. 2003. A shallow model of backchannel continuers in spoken dialogue. In *European ACL*. Citeseer, 51–58.

[4] Iwan de Kok and Dirk Heylen. 2012. Integrating backchannel prediction models into embodied conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 268–274.

[5] Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. IEEE, 168–173.

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[7] David Herrera, David Novick, Dusan Jan, and David Traum. 2011. Dialog behaviors across culture and group size. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 450–459.

[8] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech communication* 50, 6 (2008), 531–543.

[9] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 79–86.

[10] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G. Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *INTERSPEECH*. 2890–2894.

[11] Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 127–136.

[12] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* 71, 3 (2013), 250–260.

[13] Yuanchao Li, Carlos Toshinori Ishi, Koji Inoue, Shizuka Nakamura, and Tatsuya Kawahara. 2019. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Advanced Robotics* 33, 20 (2019), 1030–1041.

[14] Kohei Matsumura, Yasuyuki Sumi, and Mitsuki Sugiya. 2017. Analyzing Listeners' Empathy by Their Nonverbal Behaviors in Bibliobattle. *Journal of Information Processing* 25 (2017), 361–365.

[15] Senko K Maynard. 1989. *Japanese conversation: Self-contextualization through structure and interactional management*. Vol. 35. Praeger.

[16] 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.

[17] Toshiaki Nishio, Yuichiro Yoshikawa, Takamasa Iio, Mariko Chiba, Taichi Asami, Yoshinori Isoda, and Hiroshi Ishiguro. 2021. Actively listening twin robots for long-duration conversation with the elderly. *ROBOMECH Journal* 8, 1 (2021), 1–10.

[18] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. 2021. Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions. *Frontiers in Robotics and AI* (2021), 189.

[19] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, Jeez! or uh-huh? A listener-aware Backchannel predictor on ASR transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8064–8068.

[20] 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. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 100–108.

[21] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *CoRR* abs/2004.13637 (2020). arXiv:2004.13637 https://arxiv.org/abs/2004.13637

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

[23] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced Social Interaction with Agents*. Springer, 247–258.

[24] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. 2011. Building autonomous sensitive artificial listeners. *IEEE transactions on affective computing* 3, 2 (2011), 165–183.

[25] Khiet P Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*.

[26] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[27] Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi, Nigel G Ward, and Tatsuya Kawahara. 2016. Analysis and prediction of morphological patterns of backchannels for attentive listening agents. In *Proc. 7th International Workshop on Spoken Dialogue Systems*. 1–12.

[28] 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).