Attentive listening system with backchanneling, response generation and flexible turn-taking

Divesh Lala, Pierrick Milhorat, Koji Inoue
Masanari Ishida, Katsuya Takanashi and Tatsuya Kawahara
Graduate School of Informatics
Kyoto University
[lastname]@sap.ist.i.kyoto-u.ac.jp

Abstract

Attentive listening systems are designed to let people, especially senior people, keep talking to maintain communication ability and mental health. This paper addresses key components of an attentive listening system which encourages users to talk smoothly. First, we introduce continuous prediction of end-of-utterances and generation of backchannels, rather than generating backchannels after end-point detection of utterances. This improves subjective evaluations of backchannels. Second, we propose an effective statement response mechanism which detects focus words and responds in the form of a question or partial repeat. This can be applied to any statement. Moreover, a flexible turn-taking mechanism is designed which uses backchannels or fillers when the turn-switch is ambiguous. These techniques are integrated into a humanoid robot to conduct attentive listening. We test the feasibility of the system in a pilot experiment and show that it can produce coherent dialogues during conversation.

1 Introduction

One major application of embodied spoken dialogue systems is to improve life for elderly people by providing companionship and social interaction. Several conversational robots have been designed for this specific purpose (Heerink et al., 2008; Sabelli et al., 2011; Iwamura et al., 2011). A necessary feature of such a system is that it be an attentive listener. This means providing feedback to the user as they are talking so that they feel some sort of rapport and engagement with the system. Humans can interact with attentive listeners at any time, making them a useful tool for people such as the elderly.

Our motivation is to create a robot which can function as an attentive listener. Towards this goal, we use the autonomous android named Erica. Our long-term goal is for Erica to be able to participate in a conversation with a human user while displaying human-like speech and gesture. In this work we focus on integrating an attentive listener function into Erica and describe a new approach for this application.

The approaches to these kind of dialogue systems have focused mainly on backchanneling behavior and have been implemented in large-scale projects such as SimSensei (DeVault et al., 2014), Sensitive Artificial Listeners (Bevacqua et al., 2012) and active listening robots (Johansson et al., 2016). These systems are multimodal in nature, using human-like non-verbal behaviors to give feedback to the user. However, the backchannels are usually generated after the end of utterance and they do not necessarily create synchrony in the conversation (Kawahara et al., 2015). Moreover, the dialogue systems are still based on handcrafted keyword matching. This means that new lines of dialogue or extensions to new topics must be handcrafted, which becomes impractical.

In this paper we present an approach to attentive listening which integrates continuous backchannels with responsive dialogue to user statements to maintain the flow of conversation. We create a continuous prediction model which is perceived as being better than a model which predicts only after an IPU (inter-pausal unit) has been received from the automatic speech recognition (ASR) system. Meanwhile, the statement response system detects focus words of the user's utterance and uses them to generate responses as a wh-question or by repeating it back to the user. We also introduce a novel approach to turn-taking which uses

Proceedings of the SIGDIAL 2017 Conference, pages 127–136,
Saarbrücken, Germany, 15-17 August 2017.

backchannels and fillers to indicate confidence in taking the speaking turn.

Our approach is not limited by the topic of conversation and no prior parameters about the conversation are required so it can be applied to open domain conversation. We also do not require perfect speech recognition accuracy, which has been identified as a limitation in other attentive listening systems (Bevacqua et al., 2012). Our system runs efficiently in real-time and can be flexibly integrated into a larger architecture, which we will also demonstrate through a conversational robot.

The next section outlines the architecture of our attentive listener. In Section 3 we describe in detail the major components of the attentive listener including results of evaluation experiments. We then implement this system into Erica as a proof-of-concept in Section 4, before the conclusion of the paper. Our system is in Japanese, but English translations are used in the paper for clarity.

2 System architecture

Figure 1 summarizes the components of attentive listening and the general system architecture. Inputs to the system are prosodic features, which is calculated continuously, and ASR results from the Japanese speech recognition system Julius (Lee et al., 2001).

We implement a dialogue act tagger which classifies an utterance into questions, statements or others such as greetings. This is currently based on a support vector machine and is moving to a recurrent neural network. Questions and others are handled by a separate module which will not be explained in this paper. Statements are handled by a statement response component. The other two components in the attentive listener are a backchannel generator and a turn-taking model.

Backchannels are generated by one component, while the statement response component can generate different types of dialogue depending on the utterance of the user. As part of our NLP functionalities we have a focus word extractor trained by a conditional random field (Yoshino and Kawahara, 2015) which identifies the focus of an utterance. For example, the statement "Yesterday I ate curry." would produce a focus word of "curry". We then send this information to the statement response component which generates a question response "What kind of curry?". Further details of the technical implementation are described in the next section.

The process flow of the system is as follows. The system performs continuous backchanneling behavior while listening to the speaker. At the same time, ASR results of the user are received. When the utterance unit is detected and its dialogue act is tagged as a statement, then a response is generated and then stored. However, a response is only actually output when the system predicts an appropriate time to take the turn. This is because the user may wish to keep talking and the system should not interrupt. Thus, we can manage turn-taking more flexibly.

In summary, the three major components required for attentive listening are backchanneling, statement response and turn-taking.

3 Attentive listening components

In this section we describe the three major components of attentive listening. We evaluate each of these components individually.

3.1 Continuous backchannel generation

Our goal is to increase rapport (Huang et al., 2011) with the user by showing that the system is interested in the content of the user's speech. There have been many works on automatic backchannel generation, with most using prosodic features for either rule-based models (Ward and Tsukahara, 2000; Truong et al., 2010) or machine learning methods (Morency et al., 2008; Ozkan et al., 2010; Kawahara et al., 2015).

In this work we use a model in which backchanneling behavior occurs continuously during the speaker's turn, not only at the end of an utterance. We take a machine learning approach by implementing a logistic regression model to predict if a backchannel would occur 500ms into the future. We predict into the future rather than at the current time point, because in the real-time system Erica requires processing time to generate nodding and mouth movements that synchronize with her utterance. We trained the model using a counseling corpus. This corpus consisted of eight one-to-one counseling sessions between a counselor and a student and were transcribed according to the guidelines of the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003).

The model makes a prediction every 100ms by using windows of prosodic features of sizes 100, 200, 500, 1000 and 2000 milliseconds. For a win-

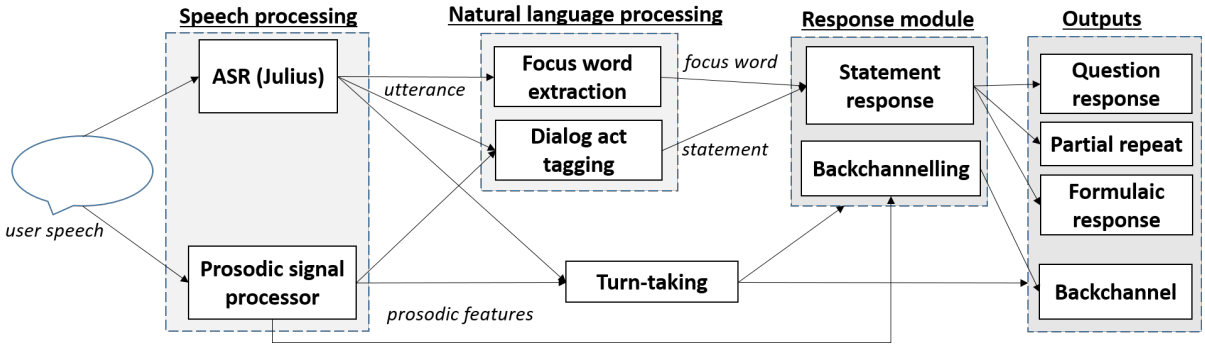

Figure 1: System architecture of attentive listener.

dow size s, feature extraction is conducted within windows every s milliseconds before the current time point, up to a maximum of $4s$ milliseconds. For example, for a time window of 100ms, prosodic features are calculated inside windows starting at 400, 300, 200 and 100 milliseconds before the current time point. The prosodic features are the mean, maximum, minimum, range and slope of the pitch and intensity. Finally, we add the durations of silence, voice activity, and overlap of the speaker and listener.

We conducted two evaluations of the backchannel timing model. The first is an objective evaluation of the precision and recall. We used 8-fold cross validation and tested on individual sessions. We compared against a baseline model which generated a backchannel after every IPU (**Fixed**) and an IPU-based model based on logistic regression which also predicted after every IPU using additional linguistic features (**IPU-based**). Our model showed that the most influential prosodic feature was the range and maximum intensity of the speech, with larger windows located just before the prediction point generally being more influential than other windows. Although we have no quantitative evidence, we propose that a reduction in the intensity of the speech provides an opportunity for the listener to produce a backchannel. The results are displayed in Table 1.

Model	AUC	Prec.	Rec.	F1
Time-based	0.851	0.344	0.889	0.496
IPU-based	0.809	0.659	0.512	0.576
Fixed	0.500	0.146	1.000	0.255

Table 1: Prediction results for backchannel timing.

We see that the time-based model performs better than the baseline and the IPU-based model with a high AUC and recall. The precision is fairly low, due to predicting a large number backchannels even though none in the corpus are found.

We also conducted a subjective evaluation of this model by comparing against the same models as the objective evaluation. We also included an additional counselor condition, in which backchannels in the real corpus were substituted with the same recorded pattern.

Participants in the experiment listened to recorded segments from the counseling corpus, lasting around 30-40 seconds each. We chose segments where the counselor acted as an attentive listener by only responding through the backchannels used in our model. The counselor's voice for backchannels was generated using a recorded pattern by a female voice actress. We created the different conditions for each recording by applying our model directly to the audio signal of the speaker. The audio channel of the counselor's voice was separated and so could be removed. When the model determined that a backchannel should be generated at a timepoint, we manually inserted the backchannel pattern into the speaker's channel using audio editing software, effectively replacing the counselor's voice.

Each condition was listened to twice by each participant through different recordings selected at random. Subjects rated each recording over five measures - naturalness and tempo of backchannels (Q1 and Q2), empathy and understanding (Q3 and Q4) and if the participant would like to talk with the counselor in the recording (Q5). Each measure was rated using a 7-point Likert scale.

For analysis we conducted a repeated measures ANOVA with Bonferroni corrections. Results are

shown in Table 2. Our proposed model outperformed the baseline models and was comparable to the counselor condition.

	Fixed	IPU	Couns.	Time-based
Q1	2.74*	3.92*	4.55	4.48
Q2	3.06*	4.05	4.86	4.61
Q3	2.44*	3.75*	4.25	4.58
Q4	2.55*	3.95	4.38	4.39
Q5	2.35*	3.64*	4.23	4.21

Table 2: Average ratings of backchannel models. Asterisks indicate the difference is statistically significant from the proposed model.

The results of both evaluations show the need for backchannel timing to be done continuously and not just at the end of utterances.

3.2 Statement response

The statement response component is triggered for statements and outputs when the system takes a turn. The purpose is to encourage the user to expand on what they have just said and extend the thread of the conversation. The statement response tries to use a question phrase which repeats a word that the user has previously said. For example, if the user says "I will go to the beach.", the statement response should generate a question such as "Which beach?". It may also repeat the focus of the utterance back to the user to encourage elaboration, such as "The beach?".

Our approach uses wh-questions as a means to continue the conversation. From a linguistic perspective, they are described in question taxonomies by Graesser et al. (1994) and Nielsen et al. (2008) as concept completions (who, what, when, where) or feature specifications (what properties does X have?). We observe that listeners in everyday conversations use such phrases to get the speaker to provide more information.

From a technical perspective, there are two processes for the system. The first process is to detect the focus word of the utterance. The second is to correctly pair this with an appropriate wh-question word to form a meaningful question. The basic wh-question words are similar for both English and Japanese.

To detect the focus word we use a conditional random field classifier in previous work which uses part-of-speech tags and a phrase-level dependency tree (Yoshino and Kawahara, 2015). The model was trained with utterances from users interacting with two different dialogue systems. This corpus was then annotated to identify the focus phrases of sentences.

We use a decision tree in Figure 2 to decide from one of four response types. If a focus phrase can be detected, we take each noun in the phrase, match them to a wh-question and select the pair with the maximum likelihood. We used an n-gram language model to compute the joint probability of the focus noun being associated with each question word. The corpus used is the Balanced Corpus of Contemporary Written Japanese, which contains 100 million words from written documents. We then consider the maximum joint probability of this noun and a question word. If this is over a threshold Tf, then a question on the focus word is generated. If no question is generated, the focus noun is repeated with a rising tone.

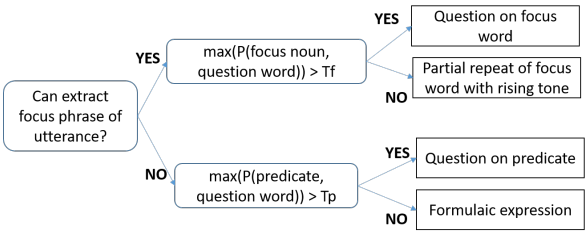

Figure 2: Decision tree of statement response system showing the four different response types.

If no focus phrase is found we match the predicate of the utterance to a question word using the same method as above. If this is above a threshold Tp, then the response is a question on the predicate, otherwise a formulaic expression is generated as a fallback response. We provide examples of each of the response types in Table 3.

We evaluated this component in two different ways. Firstly, we extracted dialogue from an existing chatting corpus created for Project Next's NLP task[1]. We selected 200 user statements from this corpus as a test set and applied the statement response system to them. Two annotators then checked if the generated responses were appropriate. The results are shown in Table 4.

The results showed that the algorithm could classify the statements reasonably well. However, in the case of a focus word being unable to be

[1] https://sites.google.com/site/dialoguebreakdowndetection/chat-dialogue-corpus

Response type	Example
Question on focus	U: Yesterday I ate **curry**. S: What kind of **curry**?
Partial repeat	U: I'll go and run a **marathon**. S: A **marathon**?
Question on predicate	U: Then I **went** out. S: Where did you **go**?
Formulaic expression	U: That's beautiful. S: Yeah.

Table 3: Examples of response types for user statements. Bold words indicate the detected focus noun or predicate of the utterance, while underlined words indicate matched question words.

Response type	Precision	Recall
Question on focus	0.63	0.46
Partial repeat	0.72	0.86
Question on predicate	0.14	0.30
Formulaic expression	0.94	0.78

Table 4: Classification accuracy of statement response system for chatting corpus.

found correctly identifying a question word for a predicate is a challenge.

Next, we evaluated our statement response system by testing if it could reduce the number of fallback responses used by the system. We conducted this experiment with 22 participants, and gathered data on their utterances during a first-time meeting with Erica. In most cases the participants asked questions that could be answered by the system, but sometimes the users said statements for which the question-answering system could not formulate a response. In these cases a generic fallback response was generated.

From the data we found that 39 out of 226 (17.2%) user utterances produced fallback responses. We processed all these utterances offline through the statement response component. From these 39 statements, 19 (47.7%) result in a statement which could be categorized into either a question on focus, partial repeat, or a question on predicate. Furthermore, the generated responses were deemed to be coherent with the correct focus and question words being applied. This would have continued the flow of conversation.

3.3 Flexible turn-taking

The goal of turn-taking is to manage the floor of the conversation. The system decides when it should take the turn using a decision model. One simple approach is to wait for a fixed duration of silence from the user before starting the speaking turn. However, we have found this is highly user-dependent and very challenging when the user continues talking. The major problem is that if the user has not finished their turn and the system begins speaking, they must then wait for the system's utterance to finish. This disrupts the flow of the conversation and makes the user frustrated. Solving this problem is not trivial so several works have attempted to develop a robust model for turn-taking (Raux and Eskenazi, 2009; Selfridge and Heeman, 2010; Ward et al., 2010).

Figure 3 displays our approach towards turn-taking behavior, rather than having to make a binary decision about whether or not to take the turn. When the user has the floor and the system receives an ASR result, our model outputs a likelihood score between 0 and 1 that the system should take the turn. The actual likelihood score determines the system's response. The system has four possible responses - silence, generate a backchannel, generate a filler or take the turn by speaking.

The novelty of our approach is that we do not have to immediately take a turn based on a hard threshold. Backchannels encourage the user to continue speaking and signal that the system will not take the turn. Fillers are known to indicate a willingness to take the turn (Clark and Tree, 2002; Ishi et al., 2006) and so are used to grab the turn from the user. However, the user may still wish to continue speaking and if they do the system won't grab the turn and so doesn't interrupt the flow of

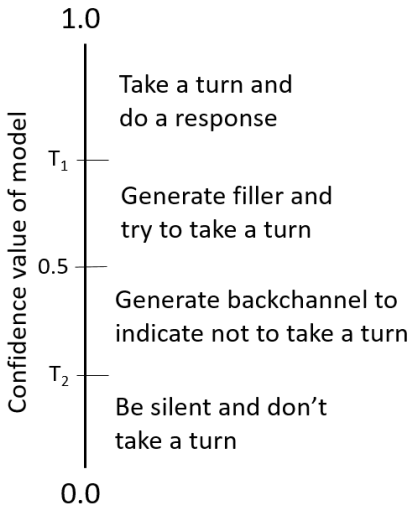

Figure 3: Conceptual diagram of Erica's turn-taking behavior. The decision of the system is dependent on the model's likelihood of the speaker has finished their turn. Decision thresholds are applied manually.

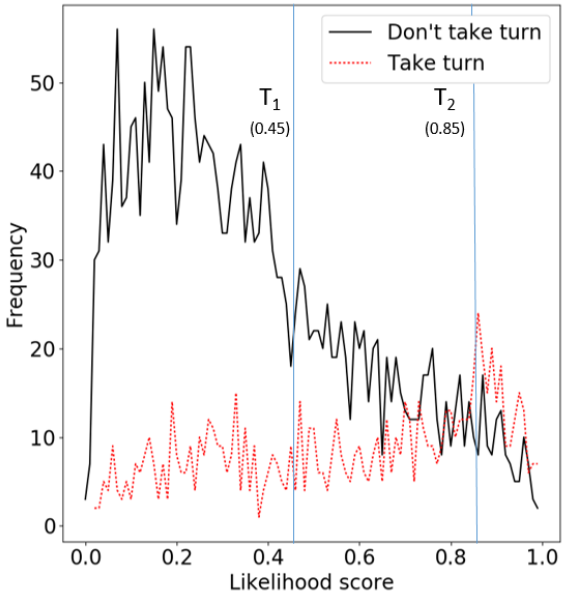

Figure 4: Distribution of likelihood scores for turn-taking.

conversation. To guarantee that Erica will eventually take the turn, we set a threshold for the user's silence time and automatically take the turn once it elapses.

To implement this system, we used a logistic regression model with the same features as our backchanneling model. We train using the same counseling corpus and features that were used for the backchanneling model. We found 25% of the outputs within the corpus to be turn changes.

Our proposed model requires two likelihood score thresholds (T_1 and T_2) to decide whether or not to be silent ($\leq T_1$) or take the turn ($\geq T_2$). We set a threshold for deciding between backchannels and fillers to 0.5. We determined T_1 to be 0.45 and T_2 to be 0.85 based on Figure 4, which displays the distributions of likelihood score for the two classes.

The performance of this model is shown in Table 5. We compared the proposed model to a logistic regression model with a single threshold at 0.5. Results are shown in Table 5.

These two thresholds degrade the recall of turn-taking ground-truth actions because the cases in between them are discarded. However we improve the precision of taking the turn, which is critical in spoken dialogue systems, from 0.428 to 0.624. The cases discarded in this stage will be recovered by uttering fillers or backchannels.

Model	Precision	Recall	F1
	3-tier		
Don't take turn	0.856	0.683	0.760
Take turn	0.624	0.231	0.337
	Binary		
Don't take turn	0.848	0.731	0.785
Take turn	0.428	0.605	0.501

Table 5: Performance of turn-taking model compared to single-threshold logistic regression.

Moreover, the ground-truth labels are based on actual turn-taking actions made by the human listener, and there should be more Transition Relevance Places (Sacks et al., 1974), where turn-taking would be allowed. This should be addressed in future work.

4 System

In this section we describe the overall system with the attentive listener being integrated into the conversational android Erica.

4.1 ERICA

Erica is an android robot that takes the appearance of a young woman. Her purpose is to use conversation to play a variety of social roles. The physical realism of Erica necessitates that her conver-

sational behaviors are also human-like. Therefore our objective is not only to undertake natural language processing, but to also address a variety of conversational phenomena.

The environment we create for Erica reduces the need to use a physical interface such as a handheld microphone or headset to have a conversation. Instead we use a spherical microphone array placed on a table between Erica and the user. A photo of this environment is shown in Figure 5.

Figure 5: Photo of user interacting with Erica.

Based on the microphone array and the Kinect sensor, we are able to reliably determine the source of speech. Erica only considers speech from a particular user and ignores unrelated noises such as ambient sounds and her own voice.

4.2 Pilot study

We conducted an initial evaluation of our system as a pilot study to demonstrate its appropriateness for attentive listening. We have observed from previous demonstrations that users often do not speak with Erica as if she is an attentive listener. Rather, they simply ask Erica questions and wait for her answers. To overcome this issue in order to evaluate the statement response system, we first provided the subjects with dialogue prompts in the form of scripts. This allowed users familiarize themselves with Erica for free conversation. Two male graduate students were subjects in the experiment and interacted with Erica in these two different tasks.

The first task was to read from four conversational scripts of 3 to 5 turns each. These scripts were not hand-crafted, but taken from a corpus of real attentive listening conversations with a Wizard-of-Oz controlled robot. Subjects were instructed to pause after each sentence in the script to wait for a statement response. If Erica replied with a question they could answer it before continuing the scripted conversation.

The second task was to speak with Erica freely while she did attentive listening. In this scenario the subjects talked freely on the subject of their favorite travel memories. They could end the conversation whenever they wished. Statistics of the subjects' turns are shown in Table 6.

	Script	**Free talk**
Turns	77	13
Avg. length per turn (sec.)	3.94	2.90
Avg. characters per turn	20.9	16.4

Table 6: Statistics for the speaking turns of the subjects.

We find that the subjects reading from the script had longer turns but the speaking rate was lower than for free talk. In other words, script reading was slower and longer. We also analyzed the distribution of response types generated from the system as shown in Table 7.

	Script	**Free talk**	**Total**
Backchannel	77	13	90
Q. on focus	14	10	24
Partial repeat	10	1	11
Q. on predicate	2	1	3
Formulaic	29	6	35
Total	132	31	163

Table 7: Distribution of response types from statement response component.

Backchannels were generated most frequently, while both questions on focus and formulaic expressions were the most common response types, with questions on focus words having the highest frequency in free conversation. Partial repeats had a much higher frequency in the scripts than in free conversation. This is because the script readings were taken from conversations which used more complex sentences than the free talk, and focus nouns for which a suitable question word could not be reliably matched.

4.3 Subjective ratings

We evaluated the system by asking 8 evaluators to listen to the recording of both the scripts and free conversation. Each evaluator was assigned

Speaker	Japanese utterance	English translation	Component
User	Kono mae, tomodachi to Awajishima ni ryokou ni ikimashita.	I once took a trip with friends to Awajishima island	
Erica	*unun*	*mhm*	Backchannel
Erica	*Doko e itta no desuka?*	*Where did you go?*	Question on predicate
User	Awajishima ni itte, sono ato bokujo nado wo-	Awajishima, then-	
Erica	*un*	*mm*	Backchannel
User	mi ni ikimashita.	went to visit a farm.	
Erica	*Doko no bokujo desu ka?*	*Where was the farm?*	Question on focus
User	Etto, namae ha chotto oboete-nain desukeredomo-	Um, I don't remember the name of it, but-	
Erica	*un*	*mm*	Backchannel
User	-ee, hitsuji toka wo mimashita.	-we saw sheep and other animals.	

Table 8: Example dialogue of user free talk conversation with attentive listening Erica.

one random script and both free conversations to evaluate. The evaluators rated each of Erica's backchannels and statement responses in terms of coherence (coherent, somewhat coherent, or incoherent) and timing (fast, appropriate, or slow). We used a majority vote to determine the overall rating of each speech act. The ratings on the coherence of each statement are shown in Figure 6.

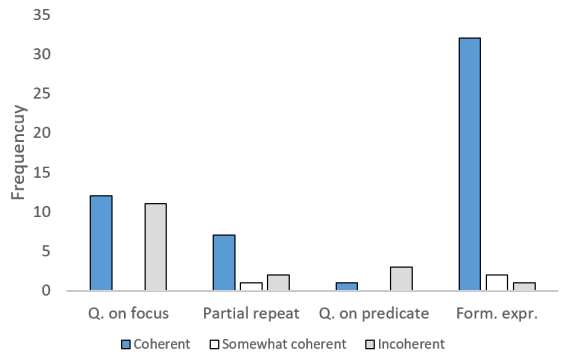

Figure 6: Rating on coherence for each response type.

We see that the results are similar to the previous evaluation of the statement response system. More than half of questions on focus words were coherent, although most of these were in response to the scripts. Formulaic expressions were mostly coherent even though they were selected at random.

Similarly, we categorized system utterances into backchannels or statements and analyzed timing. The results are shown in Figure 7.

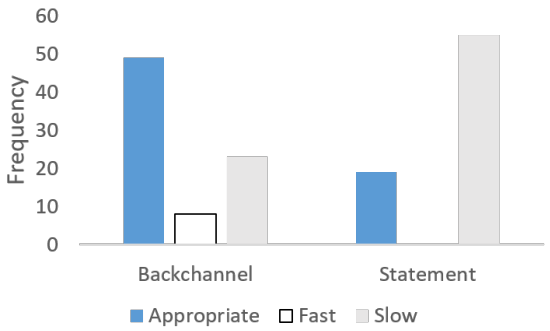

Figure 7: User rating of timing for backchannels and statements.

We can see that while most backchannels have suitable timing, statement responses are slow due to the processing of the utterance that is required.

4.4 Generated dialogue

Table 8 shows dialogue from a free talk conversation. User utterances were punctuated by backchannels and the system is able to extract a focus noun or predicate and produce a coherent response.

We also found that the system could produce a coherent response even in the case of ASR errors.

In one case the subject said "sakana tsuri wo shimashita (I went fishing.).". The ASR system generated "sakana wo sore wo sumashita", which is nonsensical. In this case, the word "fish" was successfully detected as the focus noun and a coherent response could be generated.

4.5 Analysis of incoherent statements

We also examined 17 utterances determined to be incoherent (excluding backchannels and formulaic expressions) and analyzed the reasons for these. Table 9 shows the sources of errors in the statement response with their associated frequencies.

Error source	Frequency
Incorrect question word match	5
Incoherent focus noun/predicate	4
Repeated statement	4
ASR errors	3
Focus word undetected	1

Table 9: Errors found in the generated statement responses.

Incorrect question word matching was found several times. For example, the user said "Tokyo ni ryokou ni ittekimashita (I went on a trip to Tokyo)", generating the reply "Donna Tokyo desu ka? (What kind of Tokyo?)" which does not make sense. Another source of error was the system detecting a focus noun or predicate which did not make sense. Repeated statements were also found. The subject had already explained something during the conversation but the system asked a question on it. This can be addressed by keeping a history of the dialogue. The ASR word error rate was approximately 10% for both script reading and free talk, so was not a major issue. In most cases, incorrect ASR results cannot be parsed and so a formulaic expression is produced.

4.6 Lessons from pilot study

Our pilot study showed that our system is feasible with no technical failures. Backchannels can be generated at appropriate times. Coherent responses could be generated by the system and errors in Erica's dialog can be addressed. We chose third-party evaluations for this experiment due to the small sample size and also because the subjects could not evaluate specific utterances while they were using the system.

However we intend to conduct a more comprehensive study where the subjects evaluate their own interaction with Erica. Subjects should engage in free talk, but we have found that motivating them to do so is not trivial. A reasonable metric for a full experiment is the subject's willingness to continue the interaction with with Erica which indicates engagement with the system. We can also use more objective metrics such as the number and length of turns taken by the user. Our strategy of using fillers and backchannels to regulate turn-taking should also be evaluated.

5 Conclusion and future work

In this paper we described our approach towards creating an attentive listening system which is integrated inside the android Erica. The major components are backchannel generation, statement response system, and a turn-taking model. We presented individual evaluations of each of these components and how they work together to form the attentive listening system. We also conducted a pilot study to demonstrate the feasibility of the attentive listener. We intend to conduct a full experiment with the system to discover if it is comparable to human conversational behavior. Our aim is for this system to be used in a practical setting, particularly with elderly people.

Acknowledgements

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

References

Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Maat, Gary Mckeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Etienne De Sevin, Michel Valstar, Martin Wollmer, Marc Shroder, and Bjorn Schuller. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing* 3(2):165–183.

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1):73–111.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert

Rizzo, and Louis-philippe Morency. 2014. SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support. In *International Conference on Autonomous Agents and Multi-Agent Systems*. 1, pages 1061–1068.

A. C. Graesser, C. L. McMahen, and B. K. Johnson. 1994. Question Asking and Answering. In Morton A. Gernsbacher, editor, *Handbook of Psycholinguistics*, Academic Press, pages 517–538.

Marcel Heerink, Ben Kröse, Bob Wielinga, and Vanessa Evers. 2008. Enjoyment intention to use and actual use of a conversational robot by elderly people. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. ACM, pages 113–120.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual rapport 2.0. In *International Workshop on Intelligent Virtual Agents*. Springer, pages 68–79.

Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2006. Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. In *INTERSPEECH*.

Yamato Iwamura, Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2011. Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets? In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pages 449–457.

Martin Johansson, Tatsuro Hori, Gabriel Skantze, Anja Höthker, and Joakim Gustafson. 2016. Making turn-taking decisions for an active listening robot for memory training. In Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He, editors, *Social Robotics: 8th International Conference, ICSR 2016*. Springer International Publishing, Cham, pages 940–949.

Tatsuya Kawahara, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. 2015. Toward adaptive generation of backchannels for attentive listening agents. In *International Workshop Series on Spoken Dialogue Systems Technology*. pages 1–10.

Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. 2001. Julius – an open source real-time large vocabulary recognition engine. In *EUROSPEECH*. pages 1691–1694.

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. pages 7–12.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*. Springer, pages 176–190.

Rodney D Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh, and Leysia Palen. 2008. A taxonomy of questions for question generation. In *Proceedings of the 1st Workshop on Question Generation*.

Derya Ozkan, Kenji Sagae, and Louis-Philippe Morency. 2010. Latent mixture of discriminative experts for multimodal prediction modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 860–868.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 629–637.

Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. 2011. A conversational robot in an elderly care center: an ethnographic study. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pages 37–44.

Harvey Sacks, Emanuel a Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. *Language* 50(4):696–735.

Ethan O. Selfridge and Peter A. Heeman. 2010. Importance-driven turn-bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 177–185.

Khiet P. Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Interspeech 2010*. International Speech Communication Association (ISCA), pages 3058–3061.

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

Nigel G Ward, Olac Fuentes, and Alejandro Vega. 2010. Dialog prediction for a general model of turn-taking. In *INTERSPEECH*. Citeseer, pages 2662–2665.

Koichiro Yoshino and Tatsuya Kawahara. 2015. Conversational system for information navigation based on pomdp with user focus tracking. *Computer Speech & Language* 34(1):275–291.