Chat-like Spoken Dialog System for a Multi-party Dialog Incorporating Two Agents and a User

Ryota Nishimura[1] Yuki Todo[2] Kazumasa Yamamoto[3] Seiichi Nakagawa[2]

[1] Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan
[2] Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
[3] Toyota National College of Technology, Japan

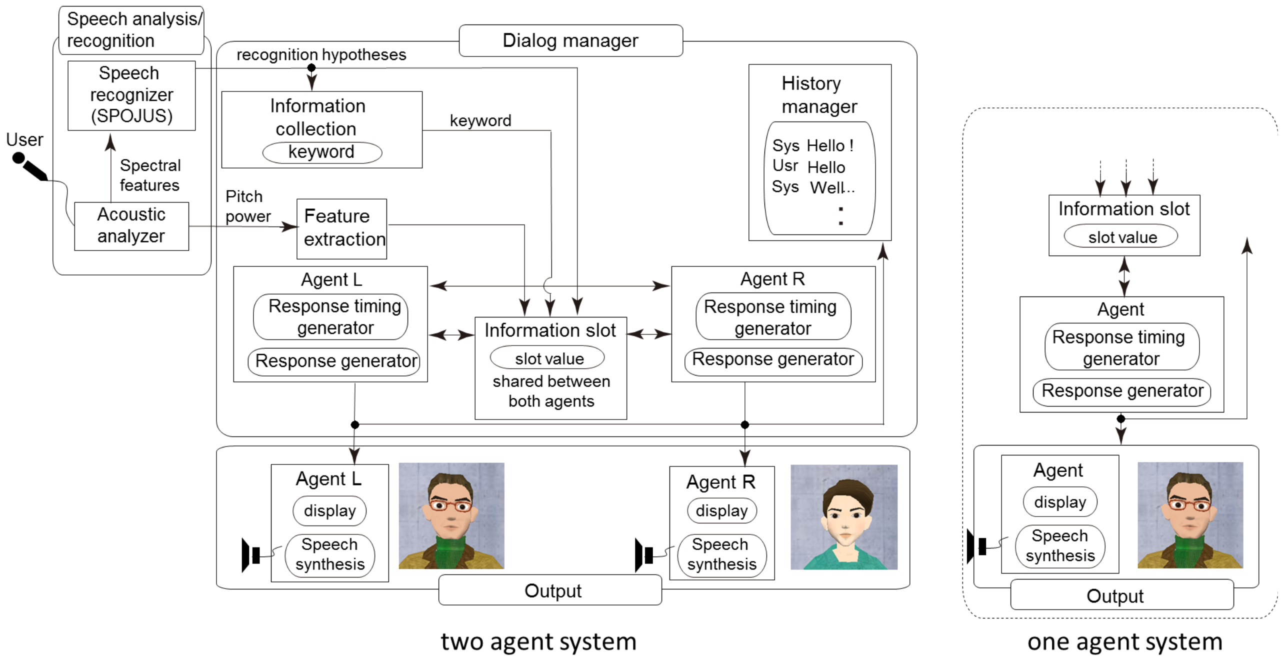
Abstract: Almost all current spoken dialog systems have treated dialog that one user talks with one agent. In this paper, to achieve the multi-party conversation (polylogue, many users/agents participate conversation), the number of system agents is increased. Three person's conversation system that treats two agents was developed. This system is extended from the spoken dialog system of two people's conversations that we have developed so far. The response timing to the user and respond type are controlled by using a decision tree. The system also reacts robustly to the user's disfluencies. The dialog tasks is "Which do you prefer, Japanese noodle (*udon*) or Chinese noodle (*ramen*) ?". We compared the three person's dialog system to the two person's system. According to the results of the experiments, the three person dialog system performed better in terms of lively conversation, and the user can talk to the agents more like chatting.

1 Introduction

Recently, the demand for speech recognition interfaces has increased and thus spoken dialog systems have been developed. Previously, we developed a spoken dialog system, which has scope for improvement in terms of achieving a more natural dialog [1][2]. Our existing dialog system mimics the interaction between human beings in spontaneous conversation and generates natural responses, including *aizuchi* (back channeling), collaborative completions, and turn-taking, whilst considering response timing. A decision tree, which refers to prosodic information and surface linguistic information as features, was employed to determine the appropriate response timings. The existing system is able to deal with *repetition*, overlap response, and barge-in. In previous research, Higashinaka et al. [3] analyzed self-disclosure and empathy in a text-based dialog system using a collected corpus and the correlation between [self-disclosure / empathy] and [closeness / satisfaction]. The dialog task was "like/dislike aspects about animals." Based on the results, a user's empathy speech represented closeness and satisfaction, while the system's empathy speech was related to that of the user. In other words, increasing the empathy of the system's speech increased user satisfaction. If a user has a good feeling about the topic, his/her self-disclosure tends to increase. In addition, Matsuzaka et al. dealt with multi-player interaction between one agent (a robot) and a multi-user scenario. Some features (such as gaze control and nodding) were introduced to ensure a natural conversation [4]. Zheng et al. developed a multi-player interactive environment in a virtual space (museum) modeled by a computer. The user was also displayed in the virtual space, and the group's behavior was modeled. Thus, they provided a multi-player dialog [5]. Based on the previous research, it is important for an interactive system to offer a sense of closeness with the system and to raise user satisfaction. We can classify spoken dialog systems into two categories; goal/task-oriented and chat-like.

In this study, we aim to develop a more enjoyable chat-like dialog system. To achieve this, we have extended our previous system[1], which allowed interaction between a single agent and the user, to handle two agents interacting with a user. In so doing we have formed a new dialog paradigm, and it is expected that the proposed system will achieve a dialog that was impossible in the previous system. Moreover, we deal with agents whose knowledge differs from hierarchical relationships. Thus, there is the possibility that by conversing with agents with different viewpoints, the user may be prompted with new ideas.

Recently, multi-party dialog has been actively studied. In the multi-party dialog between people [6, 7, 8, 9, 10] , Dielmann [11] learned a model for granting Dialog Act of multi-party dialog automatically. Shriberg et al. [12] investigated overlap/interrupt in the meeting speech data, and showed that interrupts are associated with some events (such as disfluencies) in the foreground speech. Among humans and a conversation agent [13, 14] or multi dialog agents [15, 16] , Fujie et al. conducted a real field experiment; the dialog system with a robot performed a quiz game with elderly people in an adult day-care center, and was able to become a game media which naive users such as elderly people can use and participate easily. Among humans and multi dialog agents [15, 16], in Dohsaka et al. [17], the agent decides the action depending on the situation in a multi-player conversation between humans and the conversation agents.



two agent system

one agent system

Figure 1: Schematic of the dialog system

The dialog takes place in a text-based dialog system and two users and two agents participate in the interaction. The dialog domain is a quiz, in which a question and hint are presented to the user. Two agents have the role of each setting a question and answer, and both make remarks sympathetically and egocentrically. Based on the results of the interaction experiment, the system was shown to be effective in increasing the number of user utterances and user satisfaction; in other words, the sympathetic remark from the agent improved user satisfaction, and this sparked the conversation.

Thus, the interaction of multiple agents can lead to an improvement in user satisfaction and activation of the dialog. However, as these experiments were conducted with text-based systems, the effect in spoken dialog systems remains unknown. Okamoto et al. [18] analyzed the interrelations between verbal and non-verbal modalities (speech/gaze/posture) in *manzai* (comic entertainment) dialogs. They attempted to clarify the content and timing of the appropriate behavior of the agent, and in doing so, to build a dialog system that achieved a natural conversation between agents. The reason for analyzing *manzai* dialogs was to minimize the influence of body movements, as the communication is done by verbal conversation only. From the analysis, it was found that while communicating the agent glanced towards the other agent, although his posture was directed towards his audience. The posture distribution was predominantly towards the audience for the *manzai* dialog although there was no restriction on the action. Thus, it is necessary to pay attention to posture. Okamoto et al. proposed the hypothesis that the speaker's gaze shows the "other person talking", while the posture indicates the " person addressed by the speaker". As pointed out by Okamoto et al., it is necessary to control both the posture and gaze of an agent in multi-agent dialog systems. Thus, it is necessary to display that part of a agent that meets this requirement. Based on these considerations, we have developed a spoken dialog system to handle multiple conversational agents and to increase satisfaction for the user.

2 Chat-like Spoken Dialog system

The spoken dialog system which we previously developed deals with dialog between one user and one agent. The system is now extended to the multi-party conversation, such as interaction between "two agents with different characteristics and one user". A multi-party dialog system has the following advantages:

- The conversation becomes more lively.
- Various interactive controls become possible. (All information can be shared among agents.)
- More applications of a speech dialog system can be considered.

Figure 1 shows a schematic of the dialog system for multi-party conversation with two agents. This system generates a response sentence using template matching from the result of the automatic speech recognizer (ASR). Moreover, the response type and timing are decided by inputting prosodic features into

Table 1: The difference between two agents' characters/roles

no-relationship	Agent L	Agent R
like	*Udon*	*Ramen*
dislike	*Ramen*	*Udon*
sex	male	female
hierarchical relationships	professor	assistant
a positive opinion	likes Japanese noodles	likes Japanese noodles
a negative opinion	likes Japanese noodles	likes Chinese noodles
difference in character	cheerful	quiet

the decision tree [1]. Details are given in the following paragraphs.

2.1 Domain

It is desirable to choose a conversation domain that everyone can talk freely about, and is interested in. Therefore, we chose the topic of liking/disliking two things. In the actual experiment, the topic discussed is "Which do you like, Japanese noodle (*udon*) or Chinese noodle (*ramen*) ?".

In our dialog, two agents explain/state good points and bad points, respectively, about "*udon*" and "*ramen*". In this case, it is possible to draw users into one of the opinions by ensuring that the agents have conflicting opinions. Moreover, we introduce strategies for arranging the different agents' opinions, and for drawing the user into a specific opinion.

The likes/dislikes or good/bad (positive/negative) points database has been manually constructed and includes 30 sentences.

Table 1 summaries the characteristics or roles of agents. In the case of the hierarchical relationships, the agent L is a professor and the agent R is an assistant. Although a professor has much knowledge, it is not friendly at a user. On the other hand, although an assistant has little knowledge, it is friendly at a user. And an assistant has a role which entertains between a professor and users. Thus, it is going to promote a dialog. In a positive / negative case, a positive / negative impression can also be given to a user because both agents discuss positively / negatively. And it becomes possible to draw in the target impression. If an agent's character is different, it is possible that a user does empathy to either, and changes a user's opinion, or the liveliness of a dialog changes. Thus, if interactive control is performed using two agents, various situations can be made and the possibility of interactive control will increase.

2.2 Speech analysis and recognition

The automatic speech recognizer SPOJUS [19] was employed to recognize the user input. The system have been developed by our laboratory. In the domain (Section 2.1), the number of vocabularies is about 270 words.

As acoustic features, SPOJUS uses 12 MFCCs (Mel-Frequency Cepstrum Coefficients), the first / second derivation of the MFCCs, and the first / second derivation of energy. The sampling frequency is 16 kHz. The analysis window is a Hamming window, and the frame length and frame shift are 25 ms and 10 ms, respectively. The left-to-right HMM topology has five states and four self-loops, with each state represented by four Gaussian mixtures with full covariance matrices. We used context-dependent syllable HMMs, consisting of 928 models. SPOJUS outputs the intermediate hypotheses in real-time, and it can output the recognition result less than 1∼2 seconds after finishing the utterance. The proposed system obtains the information from the intermediate hypotheses, and this is used to prepare a response, such as *repetition*.

Moreover, at the same time the system analyzes the input to extract prosodic information, such as pitch (F0) and energy, using a prosodic analyzer. These features are sent to the decision tree to produce the response timing.

2.3 Dialog management

Figure 1 gives details of the dialog manager, which consists of five sub-components ("Information collection", "Feature extraction", "Response timing generator", "Response generator", and "History manager"), and which generates response sentences using the hypotheses and prosodic information. The response timing generator, uses a decision tree to determine the response type and the timing based on the features derived from the prosodic information [1]. The pitch and energy contour patterns of the utterance are used as prosodic features. These contour patterns are expressed as regression coefficients of the F0 and log energy sequences.

The recognition results and intermediate hypotheses output by SPOJUS are sent to the information collection component, which saves the information in information slots. The slot value is sent to the response generator, which generates responses using the information. The system generates multiple patterns of responses simultaneously and the decision tree selects the most appropriate response in real-time. The selected response is sent to the output, and is presented by a speech synthesizer to the user as the response from the agent.

2.4 Output component

In the output component, each agent is displayed on separate screens by using TVML [20]. The agent's output speech is also output from two separate loudspeakers and we use a text to speech synthesized voice

Table 2: Examples of slot and values

Slot name	examples of values
user's favorite food	*udon*
user's favorite type	*miso*
user's favorite topping	deep-fried *tofu*
reason (like)	delicious
reason (disliked anti-food)	unhealthy

(GalateaTalk [21]). In the speech synthesis, there is a delay of about 500 ms. To avoid this delay, the system response is prepared (recorded) to a file beforehand (about 400 utterances) and the speech file is played when the system responds. The three person dialog system consists of male and female agents, the two person dialog system's agent consists of a male agent only.

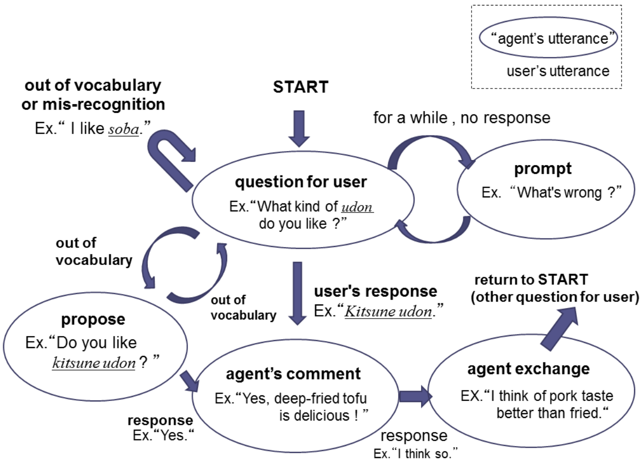


Figure 2: *State transitions in a three person dialog.*

3 Detail of dialog management

3.1 Feature extraction [1]

Here, the prosodic features used as input into the decision tree to decide the response timing and the response type are calculated based on the output of the speech analyzer. We used the first-order regression coefficients of the pitch and energy sequences in the last three regions of utterances obtained from a 55 ms length sliding window with 30 ms overlap (where the total length is 105 ms). A longer region also includes information that triggers responses, so the pitch/energy contours in the last 500 ms were also used. To describe these patterns, we adopted the first-order regression coefficients for 100 ms length segments with no overlap. The coefficients of five continuous segments describe the pattern. As these coefficients can be calculated with very little computational cost, the calculation can be done in real-time. Because real-time processing is important in spoken dialog systems, these are suitable features.

3.2 Information collection

The necessary information is extracted from the ASR result and stored in the slot. The word tag is given to the speech recognition result at each word. For example, "I[subj] like[verb] udon[food]." The word inputted into a slot is determined from this tag information. Since the word tag is registered in the speech recognition dictionary, it does not need to process a morphological analysis. The slot value is used for response generation which is possible to consider the context. Here, the conversation domain is "*udon* and *ramen.* " Therefore, examples of values stored in the slot are shown in Table 2.

3.3 Response generator

Template matching is used to generate responses in the proposed system. By comparing the speech recognition result with the response templates, a response sentence is prepared based on the matched one. Furthermore, a response sentence that considers the dialog context can be generated by using slot information.

Fig. 2 shows the state transition of the three person spoken dialog system with two agents used in this study. Speech production is carried out in the system according to the state transitions. In the figure, encircled utterances denote utterances by agents, while those depicted without circles denote user utterances. In our system, the dialog begins with a question posed to the user in the start state, "question for user". If the system does not receive any response from the user, it prompts the user to respond. If the user's utterance contains unknown words or does not match a rule defined by the system, the agent provides an example that the user can talk about. If the utterance matches a rule, the agent comments on the utterance, and the system then switches between the current agent and the other one. After the change, the dialog state returns to the start state and the dialog is repeated.

The following is an example of a dialog with two agents (Agent L and Agent R).

```
Agent L: Which do you prefer,
           udon or ramen?
User   : Well, I like ramen.
Agent L: Oh, me too.
          What kind of ramen do you like?
User   : I like miso ramen.
Agent L: I see. Miso is very delicious.
Agent R: I like udon. What do you think?
User   : I also like udon.
Agent R: I see.
```

3.4 Response timing generation

Previously, we proposed a decision tree-based response timing generator, but this was only able to produce a response after detecting the pause (at the end of the

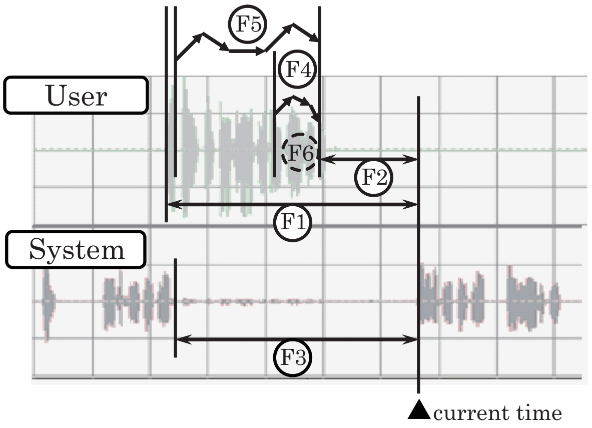


Figure 3: Features used by the decision tree

user utterance). We have modified this method to enable it to generate overlapping responses by scanning all segments (each segment length is 100 ms) continuously while the user is speaking. The response timing generator generates the response timing using the prosodic features in the decision tree. At the same time, the appropriate response is selected from the responses of the response generator.

The following features are used in determining the response timing.

F1: Duration from the start of the user's last utterance
F2: Elapsed time from the end of the previous user utterance
F3: Elapsed time from the end of the previous system utterance
F4: Pitch/energy contour of the last 100 ms (consisting of three values)
F5: Pitch/energy contour of the last 500 ms (consisting of five values)
F6: Attribute of the last word in the last recognition result (or current intermediate hypothesis)

The relation of each feature is shown in Figure 3.

Information on whether or not the response contents are prepared by the response generator is also used as a feature. Features are input into the decision tree every 100 ms. The decision tree selects the dialog action to be carried out by the system at every instance. The frequency of the responses, with the exception of *aizuchi* and *repetition*, is limited to one per user utterance. Because the system always gives an ordinary response to a user utterance, *aizuchi* and *repetition* can be used over and over as a response to the utterance.

The RWC corpus [22] was used to train the decision tree for *aizuchi*, turn-taking, and *wait*(no response). It has 48 conversations each about 10-minutes long, giving a total of 6.5 hours of dialog. The corpus consists of 16,399 utterances, covering two conversation areas: 'car sales' and 'overseas trip planning'. The speaker on one side is a professional salesperson, while the questioner / customer on the other is one of 12 non-professional men and women. C4.5 [23] was used to construct the decision tree.

3.5 History manager

The conversation history is preserved so that it can be referred to. As a result, a conversation strategy that considers the dialog context can be implemented. This component is not currently used in the system. In the future, the system will be able to communicate using the conversation history.

4 Construction of a two person dialog system from a three person dialog system

Two person dialog system (one user and one agent) was built for comparison with a three person dialog system. The two person dialog system uses the same speech recognizer, grammar, vocabulary, and templates as the three person system.

In the three person dialog system, each agent recommends his/her favorite food to the user. On the other hand, in the two person dialog system, agent recommend both foods to user.

The following is an example of a dialog with only one agent system.

```
Agent : Which do you prefer,
         udon or ramen?
User  : Well, I like ramen.
Anget : Oh, I like both.
        What kind of ramen do you like?
User  : I like miso ramen.
Agent : I see. Miso ramen is very delicious.
Agent : I think miso udon is also delicious.
User  : You're right.
Agent : What do you think about udon?
```

5 Experimental results

5.1 Setup

Subjects in the experiment consisted of 20 males in their twenties. Each subject evaluated both the three person and two person dialog systems by interacting with them. Two person dialog system (one user and one agent) was built for comparison with a three person dialog system. The two person dialog system uses the same speech recognizer, grammar, vocabulary, and templates as the three person system. In the three person dialog system, each agent recommends his/her favorite food to the user. On the other hand, in the two person dialog system, agent recommend both foods to user.

Subjects first viewed a video about the systems, and then used the dialog systems for a few minutes to become familiar with how to use them. We told the subjects that they had to talk with agents as long

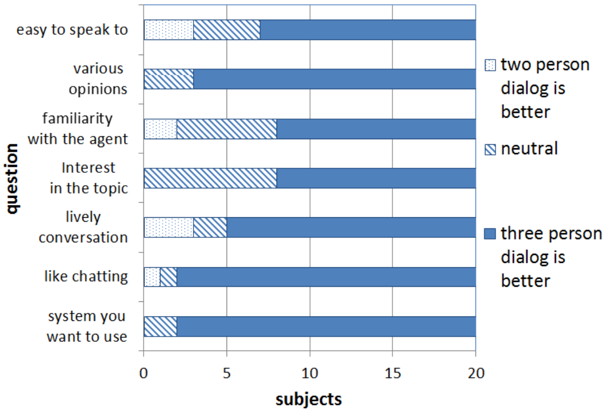


Figure 4: *Relative evaluation: "Two person dialog is better" represents those who gave a 1 or 2 point answer, while "three person dialog is better" represents those who gave a 4 or 5 point answer to the question. Neutral subjects were those who gave a 3 as their answer to a question.*

as possible until we signaled. Thereafter, each subject interacted with both dialog systems for about 5 minutes, and then stopped talking. After using both systems, subjects completed a survey questionnaire. Half the subjects used the two systems in reverse order. The questionnaire included the following questions:

1. Which system is easier to interact with? (two person dialog (1 2 3 4 5) three person dialog)
2. In which system did you obtain various opinions from the agent(s)?
3. In which system did you feel familiarity with the agent(s)?
4. Which system's topic (*udon* and *ramen*) was of interest to you?
5. In which system did you have a lively conversation with the agent(s)?
6. With which system did you prefer chatting?
7. Which system would you want to use again if the content and timing of its responses were more natural?

5.2 Subjective evaluation

5.2.1 Relative evaluation

Answers to the survey questions are summarized in Fig. 4. Based on the answers to questions 2 and 5–7, most subjects preferred the three person dialog system. With regard to all questions, many subjects preferred the three person dialog system significantly (T-test, two-sided, $p < 0.05$). With regard to Q2, an example response was: "From talking by three persons, it is thought that more opinions come out. " With regard to Q5, the user said "I felt that the system had many topics." In practice, there was no difference in the quantity of topic. With regard to Q6, an example response was: "the conversation with the two person dialog system feels like a question-answering system". With regard to Q7, 80% (or more) of subjects gave highly rating to the three person system. There was such an opinion: "There are two agents, and if they have different opinion each other, a dialog will more lively. "

On the other hand, there were no significant differences in Q3 and Q4. However, there was also a useful opinion. Regarding familiarity with the agents (Q3), the subjects were more familiar in the three dialog system as "the roles of the agents were clear in the three person dialog system". With regards interest in the topic (Q4), the subjects were of the opinion that "We got useful negative feedback from the agents in the three person dialog system".

However, with regard to questions 1 and 5, the opinions of the subjects were split. Those subjects who gave a high evaluation to the three person dialog system were of the opinion that "it was not easy to talk to the two dialog system" and "I could talk to the agents in the three person dialog system with intent". Conversely, subjects who gave a high rating to the two person dialog system said that "it felt like I was facing a barrage of questions from the agents in the three person dialog system" and "I had to wait for the end of conversations between two agents in the three person dialog system". In Q1 and Q5, there was a subject with high rating of two person dialog because timing control between agents had not considered. In the experiment, we used the fixed value for response timing. In a future work, we intend to control the timing of the conversation between the agents as well.

In addition, there was a high correlation (0.45) between Q5 and Q7. From this fact, we guess that the users want to use a system that can lively interact.

5.2.2 Absolute evaluation

In addition to the relative evaluation, each subject evaluated the two and three person dialog systems using an absolute evaluation scale ranging from (disagree) 1–5 (agree) for questions such as " Is it easy to talk to the agent(s)?" Answers to the survey questions are given in Fig. 5. Responses to all the questions with respect to the three person dialog system were rated more highly than those for the two person dialog system, especially the evaluation of "easy to speak to"(T-test, $p < 0.1$), "various opinions", "lively conversation" and "like chatting"(each $p < 0.05$). Thus, the results of the experiments show that the three person dialog system was rated more highly in terms of ease of conversation and users can talk with the agents more like chatting.

5.3 Objective evaluation

As an objective evaluation, Table 3 shows a part of the automatic speech recognition (ASR) performance (Cor), Out Of Vocabulary rate (OOV), and frequency of dialog phenomena, that is, for only typical 9 speaker

Table 3: Speech recognition performance and frequency of dialog phenomena in two and three person systems. (Speakers 1–4 have the best 4 ASR performance (Cor) and speakers 17–20 have the worst 4 ASR performance.)

speaker	Correct [%]		OOV [%]		dialog duration		# user turns		# system turns	
	two	three	two	three	two	three	two	three	two	three
1	72.5	82.8	4.5	0.0	4’21”	4’14”	38	35	62	52
2	73.5	81.3	2.9	4.5	4’18”	4’50”	44	46	59	63
3	80.7	73.6	2.1	4.9	4’23”	4’32”	44	45	62	60
4	70.4	76.2	2.4	5.4	4’48”	5’03”	66	52	79	72
17	49.0	54.1	2.1	1.3	5’42”	6’00”	49	52	70	76
18	49.4	44.0	10.0	9.5	5’11”	5’30”	66	66	82	78
19	45.3	44.1	10.3	7.1	4’43”	4’48”	59	56	81	77
20	**55.4**	**27.9**	**7.7**	**17.7**	5’58”	5’50”	48	48	67	63
average	62.7	61.3	4.6	6.3	4’56”	5’04”	50.2	48.0	70.0	69.4
correlation with Correct			-0.46	-0.65	—	—	-0.22	-0.40	—	—

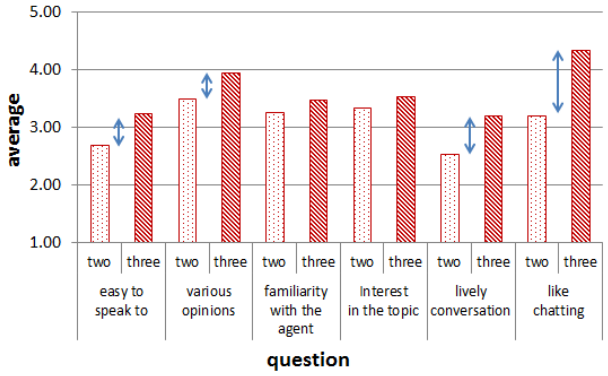


Figure 5: Absolute evaluation: average

(users) out of 20 speakers. Speakers 1–4 have the best 4 ASR performance (Cor) and speakers 17–20 have the worst 4 ASR performance. Included in the system’s turn is *aizuchi*. All the dialogs comprised about 100 turns over five minutes. Regarding the correlation between ASR performance (word correct rate) and the OOV (two, three) indicates a significant correlation.

Moreover, speakers 7 and 20 gave higher scores to the two person dialog system in the relative evaluation. However, according to the table, the system had many turns in the three person dialog with speaker 7, and as a result, in his evaluation, he stated that it was not easy to talk to the agents. Moreover, speaker 20 had a much lower ASR performance in the three person dialog than in the two person dialog. Thus, if ASR performance and the frequency of the system’s response worked better, we could conclude that users had an overall good impression of the three person dialog system. Interestingly, in all speakers, regarding the correlation between Cor (ASR) performance and “like chatting” indicates a significant correlation 0.40 in the two person dialog system in absolute evaluation and 0.13 in the three person dialog system. On the other hand, “like chatting” of absolute evaluation is a higher evalutaion in the three person dialog system than the two person dialog system as shown in Fig. 5. So, the subjects felt like that the conversation with the three person dialog system is chat, independent of ASR’ performance.

6 Conclusion

In this paper, a spoken dialog system consisting of one user and one agent was extended to a three person conversation system with two agents. Both systems were compared in terms of user behavior and satisfaction. Based on the results of the experiments, the three person dialog system achieved better results in terms of “familiarity with the agent”, “interest in the topic”, especially, “easy to speak to” , “various opinion”, “lively conversation” and “like chatting”.

In future work, we intend to compare both systems in another domain (e.g., trip to Hokkaido (snowy region) vs. trip to Okinawa (tropical region)) and to compare synthesized speech with recorded voice with regard to the response speech. Since, the fixed value was used for the utterance timing in agents’ dialog in the experimet, we would like to also control this timing.

References

[1] Nishimura, R. and Nakagawa, S.: Response timing generation and response type selection for a spontaneous spoken dialog system, *Proceedings of 2009 IEEE Workshop on Automatic Speech Recognition and Understanding(ASRU-2009)*, pp. 462–467 (2009).

[2] Itoh, T., Kitaoka, N. and Nishimura, R.: Subjective experiments on influence of response timing in spoken dialogues, *Proceedings of the Interspeech 2009*, pp. 1835–1838 (2009).

[3] Higashinaka, R., Dohsaka, K. and Isozaki, H.: Effects of Self-Disclosure and Empathy in

Human-Computer Dialogue, *2008 IEEE Workshop on Spoken Language Technology (SLT 2008)*, pp. 109–112 (2008).

[4] Matsusaka, Y., Tojo, T. and Kobayashi, T.: Conversation Robot Participating in Group Conversation(Special Issue on the 2001 IEICE Excellent Paper Award), *IEICE transactions on information and systems*, Vol. 86, No. 1, pp. 26–36 (2003).

[5] Jun, Z., Xiang, Y. and San, C. Y.: Designing multiparty interaction support in Elva, an embodied tour guide, *AAMAS '05: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 929–936 (2005).

[6] Kathol, A. and Tur, G.: Extracting Question/Answer Pairs in Multi-party Meetings, *Proceedings of ICASSP 2008*, pp. 5053–5056 (2008).

[7] Ba, S. O. and Odobez, J.-M.: Multi-party Focus of Attention Recognition in Meetings From Head Pose and Multimodal Contextual Cues, *Proceedings of ICASSP 2008*, pp. 2221–2224 (2008).

[8] Hakkani-Tur, D.: Towards Automatic Argument Diagramming of Multiparty Meetings, *Proceedings of ICASSP 2009*, pp. 4753–4756 (2009).

[9] Marin, A., Wu, W., Zhang, B. and Ostendorf, M.: Detecting Targets of Alignment Moves in Multiparty Discussions, *Proceedings of ICASSP 2012*, pp. 5129–5132 (2009).

[10] Mayfield, E., Adamson, D. and Rosé, C. P.: Hierarchical Conversation Structure Prediction in Multi-Party Chat, *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 60–69 (2012).

[11] Dielmann: DBN Based Joint Dialogue Act Recognition of Multiparty Meetings, Proceedings of ICASSP ' 07, pp. 133–136 (2007).

[12] Shriberg, E., Stolcke, A. and Baron, D.: Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation, *Proceedings of the Interspeech 2009*, pp. 1359–1362 (2009).

[13] Klotz, D., Wienke, J., Peltason, J., Wrede, B., Wrede, S., Khalidov, V. and Odobez, J.-M.: Engagement-based Multi-party Dialog with a Humanoid Robot, *Proceedings of the SIGDIAL 2011*, pp. 341–343 (2011).

[14] Fujie, S., Matsuyama, Y., Taniyama, H. and Kobayashi, T.: Conversation Robot Participating in and Activating a Group Communication, *Proceedings of the Interspeech 2009*, pp. 264–267 (2009).

[15] Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S. and White, K.: Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides, *Proceedings of the 10th international conference on Intelligent virtual agents*, pp. 286–300 (2010).

[16] Traum, D., Marsella, S. C., Gratch, J., Lee, J. and Hartholt, A.: Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents, *Proceedings of the 8th international conference on Intelligent Virtual Agents*, pp. 117–130 (2008).

[17] Dohsaka, K., Asai, R., Higashinaka, R., Minami, Y. and Maeda, E.: Effects of Conversational Agents on Human Communication in Thought-Evoking Multi-Party Dialogues, *SIGDIAL*, pp. 217–224 (2009).

[18] Okamoto, M., Ohba, M., Enomoto, M. and Iida, H.: Multimodal Analysis of Manzai Dialogue Toward Constructing a Dialogue-Based Instructional Agent Model, *journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 20, No. 4, pp. 526–539 (2008).

[19] Kai, A. and Nakagawa, S.: A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar, *ICSLP-92*, pp. 257–260 (1992).

[20] M.Hayashi: TVML(TV program Making Language) Make Your Own TV Programs on a PC!, *International Conferences, Virtual Studios And Virtual Production* (2000).

[21] Kawamoto, S., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., Morishima, S., Yotsukura, T., Kai, A., Lee, A., Yamashita, Y., Kobayashi, T., Tokuda, K., Hirose, K., Minematsu, N., Yamada, A., Den, Y., Utsuro, T. and Sagayama, S.: Open-source software for developing anthropomorphic spoken dialog agent, *Proc. of PRICAI-02, International Workshop on Lifelike Animated Agents*, pp. 64–69 (2002).

[22] Tanaka, K., Hayamizu, S., Yamasita, Y., Shikano, K., Itahashi, S. and Oka, R.: Design and Data Collection for a Spoken Dialogue Database in the Real World Computing Program, *Proc. ASA-ASJ Third Joint Meeting*, pp. 1027 –1030 (1996).

[23] J. Quinlan, R.: C4.5:Programs for machine learning, *Morgan Kaufmann* (1992).