

Response Timing Generation and Response Type Selection for a Spontaneous Spoken Dialog System

Ryota Nishimura #¹, Seiichi Nakagawa #²

Department of Information and Computer Sciences, Toyohashi University of Technology,
1-1, Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan

¹nishimura@slp.ics.tut.ac.jp

²nakagawa@slp.ics.tut.ac.jp

Abstract—If a dialog system can respond to a user as naturally as a human, the interaction will appear smoother. In this research, we aim to develop a dialog system that emulates human behavior in a chat-like dialog. The proposed system makes use of a decision tree to generate chat-like responses at the appropriate times. These responses include “*aizuchi*” (back-channel), “repetition”, “collaborative completion”, etc. The system also reacts robustly to the user’s overlapping utterances (barge-in) and disfluencies. The subjective evaluation shows that there is a high degree of naturalness in the timing of ordinary responses, overlap, and *aizuchi*, and that the dialog system exhibits user-friendly behavior. The recorded voices system was preferred, and almost all subjects felt familiarity with *aizuchi*, and the barge-in was also useful.

I. INTRODUCTION

Recently, there has been increased interest in and demand for interfaces using ASR (Automatic Speech Recognition). Because traditional systems provide no reaction to user utterances, a user cannot distinguish whether the system has recognized the utterance correctly. Despite several systems having been developed thus far, spoken dialog systems still tend to give a *stiff* impression.

In Japanese human-to-human dialog, well-timed responses such as *aizuchi* (sometimes called ‘back-channel’) and turn-taking ensure a smooth dialog. The properties of *aizuchi* and turn-taking indicate that pitch (F0) and power are mainly related to generating *aizuchi* and turn-taking. Various real-time *aizuchi* generation systems have been developed [1] that use pitch (i.e., the inverse of the fundamental frequency (F0)) and pause duration as features. Some natural turn-taking timing detection systems have also been developed [2]. Although various kinds of responses need to be considered to emulate human responses, these previous studies dealt only with an individual kind of response.

The purpose of this study is to generate natural responses, including *aizuchi*, collaborative completions, and turn-taking whilst considering response timing. In this paper, we deal with the chat-like conversation that means non-task-oriented and mixed-initiative dialogs, and the dialog through full-duplex communications. A decision tree that contains referring prosodic information and surface linguistic information as features is employed to decide the appropriate response timing. Using this timing generation method, a human-friendly spoken dialog system has been developed [3]. The proposed system is able to deal with *repetition*, overlap response and

barge-in, whereas the system in [3] does not consider these phenomena. One of our system’s goals is to provide a user-friendly interface, so that humans will want to interact with it.

In this paper, we first discuss the implementation of the response timing generator for the dialog system. Thereafter, we present the results of the subjective evaluation of the dialog system.

II. MODELING RESPONSE TIMING

The decision of response selection and the response timing was modeled by using the human-to-human conversation corpus. To model these response behavior, a decision tree was used, and the model was implemented to the system. To imitate the human-to-human conversation, this system treats *aizuchi*, repetition, collaborative completion and other ordinary responses. *Aizuchi* was the signal to make hearers continue to speak, to indicate his/her understanding of the utterances, and to express assent to them. The repetition was used to confirm the content while talking. Collaborative completion was used by that the hearer often overlaps speaker’s utterances with same contents or sometimes complements the speaker’s utterance by the prediction of the latter half of the utterance from the first half so as to complete an utterance. Moreover, the overlaps as response timing is also treated. In human-to-human conversation, there is a correlation between the familiarity of conversation and the overlap frequency in the conversation [4]. To achieve a familiar conversation, it is important to treat the overlap.

A. Features of response timing generation

According to [5] and [6], the contour patterns of pitch and power are related to the timing of response generation. For example, when the pitch and/or power contours of the mora at the end of an utterance follow various particular patterns, the conversational partner’s *aizuchi* or turn-taking is triggered. Thus, we used the first-order regression coefficients of the pitch and power sequences in the last three regions of utterances obtained from a 55-ms 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/power 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

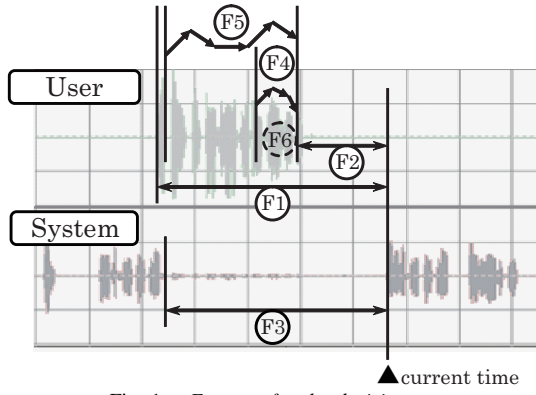


Fig. 1. Features for the decision tree.

coefficients of five continuous segments express the pattern. As these coefficients can be calculated with very little computational cost, the calculation can be done in real-time.

The repetition and collaborative completion occur when a keyword of the conversation topic is input by the user. When a speaker feels that the listener is unable to keep up with the conversation (for example, when giving a telephone number), the speaker divides the utterance into several ‘fragments’. In such cases, the listener often uses repetitions of the fragments (or keywords) to indicate the status of his/her understanding. To imitate this behavior, the response generator should detect keywords in user utterances.

The following features are used in implementing the above for the decision tree [7].

- Duration from the start of the user utterance (F1)
- Elapsed time from the end of the user utterance (F2)
- Elapsed time from the end of the system utterance (F3)
- Pitch/power contour of the last 100 ms (F4)
- Pitch/power contour of the last 500 ms (F5)
- Attribute of the last word in the last recognition results (or current intermediate hypotheses) (F6)

Figure 1 illustrates these features for the decision tree which decides the response timing.

B. Response timing generation using decision tree

Previously, we proposed a decision tree-based response timing generator [3], but this was only able to produce a response after detecting the pause (at the end of the user utterances). We have modified this method to enable it to generate overlapping responses by scanning all segments whenever the user speaks. A part of the decision tree is shown in Figure 2. The correspondence of the terms is shown in Table I.

The response timing generator decides the response timing as well as the selection of response sentence from responses prepared by the response generator, using a decision tree based on the features introduced in Section II-A. Information on whether or not the response contents were 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 a dialog act for the system to do at every instance, from *aizuchi*, repetition, collaborative completion, ordinary response, and

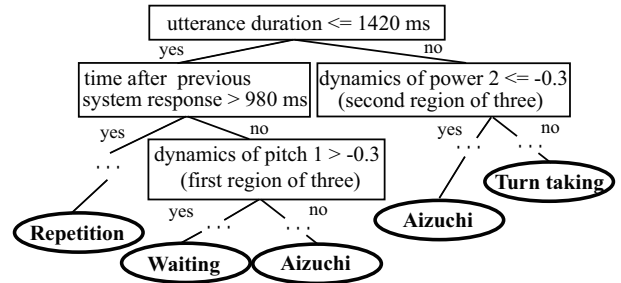


Fig. 2. Part of the decision tree.

TABLE I

CORRESPONDENCE OF THE TERMS.

response type (dialog act)	decision tree output (dialog act + wait)	response timing
<ul style="list-style-type: none"> • aizuchi • repetition • collaborative completion • ordinary response 	<ul style="list-style-type: none"> • aizuchi • repetition • collaborative completion • ordinary response • wait 	<ul style="list-style-type: none"> • overlap • (turn-taking) • (wait)

wait, as illustrated in Figure 3. *Wait* means “do not output any response”. The frequency of the responses, except *aizuchi* and repetition, is limited to one for each user utterance. Because the system gives an ordinary response for a user utterance, on the other hand, *aizuchi* and repetition can be used many times as the response to a user utterance. There are four types of responses, i.e. *aizuchi*, repetition, collaborative completion, and ordinary response. Collaborative completion and ordinary response are considered as turn-taking. The overlap indicates a timing of these response, so the case such as “repetition and overlap” exists.

The RWC corpus [8] was used to train the decision tree for *aizuchi*, turn-taking, and *wait*. The RWC has 48 conversations each about 10-minutes long, giving a total of 6.5 hours. 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, and the questioner / customer on the other side is one of 12 non-professional men and women. C4.5 [9] was used to construct the decision tree.

Based on the decision tree, *aizuchi* occurs when two seconds or more have elapsed from the latest response of the system, when the 0.5 seconds or more have elapsed from the start of user utterance, and when the pitch contour is flat or High-Low slope. With regards the other phenomena, namely repetition and collaborative completion, there were not enough training data in the corpus. The frequency of repetition and collaborative completion is less than 1%. However, these

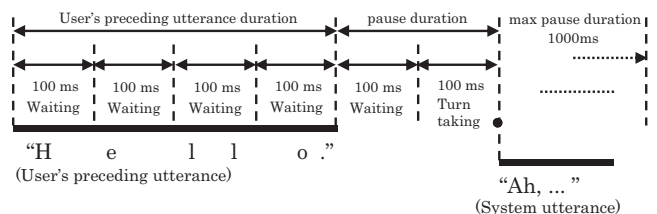


Fig. 3. Response timing generated by the decision tree

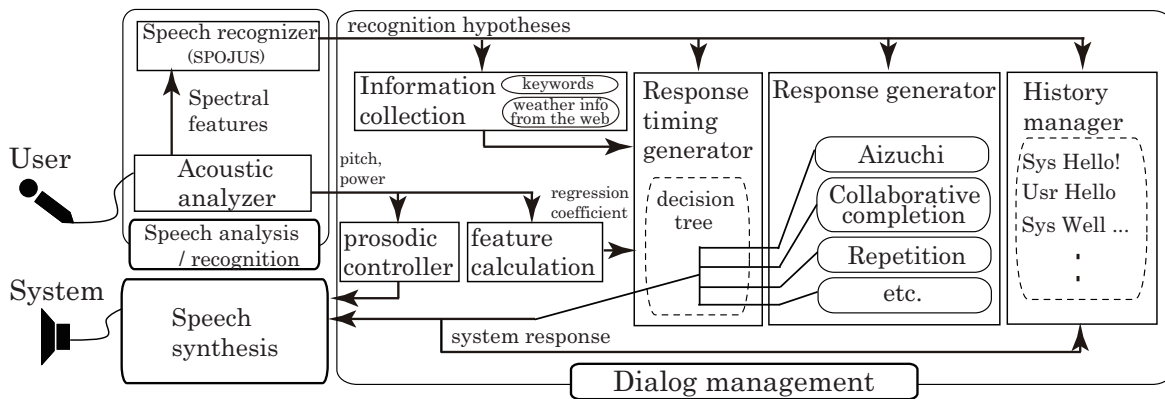


Fig. 4. Schematic diagram of the dialog system.

phenomena exist in the human-to-human conversation, and the implement of these phenomena is important for imitating human-to-human conversation. Therefore, we added some rules manually.

Based on the hand-crafted rules, repetition occurs when two seconds or more have elapsed from the latest response of the system, and when the last word in the recognition hypothesis is a city name in our dialog domain. Collaborative completion occurs immediately when the response was prepared (i.e. the user input matches with the template and when the system response is prepared).

Furthermore, the system has an exceptional rule to continue the dialog, such that the system prompts the user to say something after a long pause (6 seconds in our system). Furthermore, when a pause of over 1000 ms occurs after the last user utterance, the system gives one of possible responses to the user without consulting the decision tree.

III. DIALOG SYSTEM

Figure 4 shows the novel architecture of the proposed spoken dialog system that can deal with the various phenomena described above. For the pilot study, we chose weather information as the dialog domain, for the following three reasons. First many subjects can talk comfortably about this domain, and secondly it can deal with real information from the WEB. The final reason is that the domain and utterances can be limited naturally.

A. Speech analysis and recognition

The speech recognizer SPOJUS [10], [11] was employed to recognize the user input. There are two versions of SPOJUS: an n-gram based large vocabulary continuous speech recognizer, and a CFG (Context Free Grammar) based one. We used the latter in our system.

SPOJUS uses 12 MFCCs (Mel-Frequency Cepstrum Coefficients), the first/second derivation of the MFCCs and the first/second derivation of power as acoustic features. 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 HMM topology has four states and five 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. The proposed system obtains the information from the intermediate hypotheses, and this is used to prepare a response such as *repetition*.

SPOJUS used a vocabulary of 300 words including city names, dates, types of weather, fillers, etc., together with word class information. In the recognition results (or intermediate hypotheses), an attribute is attached to each word, and this information is useful for detecting keywords. For example, “How[*Question*] is[] the[] weather[*Weather*] in Hamamatsu[*city name*] today[*Date*]”. Our system focuses on weather information, and thus the attributes of the keywords include place-names, dates, weather in topical places, etc. The attribute of the last word in the hypotheses (or intermediate hypotheses) is used as a feature.

Simultaneously, the system analyzes the input to extract prosodic information, such as pitch (F0) and power, using a prosodic analyzer.

B. Dialog management

Details of the dialog manager are illustrated in Figure 4. It is composed of six sub-components and generates response sentences using the hypotheses and prosodic information. One of the six sub-components uses a decision tree to determine the timing based on the features derived from the prosodic information. The pitch and power contour patterns of the utterance are used as prosodic features. These contour patterns are expressed by regression coefficients of the F0 and log power sequences. The pitch and power are used to control the prosodic information of the system output responses.

Recognition results and intermediate hypotheses output by SPOJUS are sent to the information collection component. Then, the system saves the information into information slots. The slot information is sent to the response generator, which generates responses using the information. The system generates multiple patterns of responses simultaneously, and the decision tree then selects the most appropriate response from these in real-time.

The response generator prepares response sentences using an ELIZA[12]-like procedure with slot-based history management, in addition to the recognition hypotheses. Thus, the

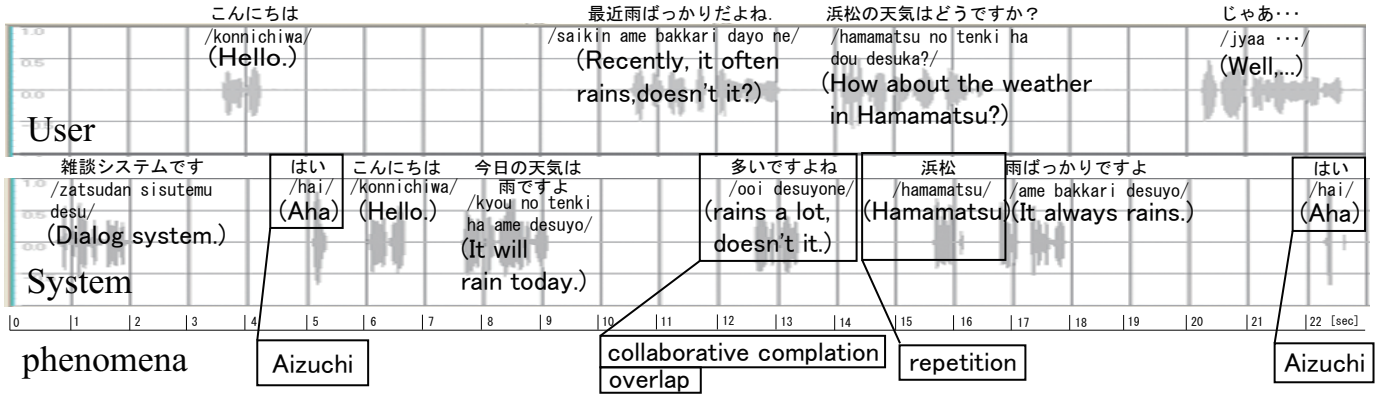


Fig. 5. Example of a dialog between a user and the system.

response generator also serves as a simple dialog manager.

C. Prosodic controller

The prosodic controller generates prosodic information for the system output responses, such as pitch, power, and speech rate. The prosodic controller controls the mean values of the prosodic information for the entire utterance but does not control the change in prosodic pattern within a single utterance. The prosodic information is sent to the speech synthesizer. Details of the prosodic controller can be found in [4].

D. Speech synthesizer

To output responses as speech, we use a recorded human voice or text to speech synthesizer voice. A female voice was used to record 3410 sentences, including *greetings*, *aizuchi*, and weather information, with a familiar and lively intonation. GalateaTalk [13] is used as the speech synthesizer, which can control speaker type, voice tone (intonation), and speech rate. Same sentences are prepared for both recorded human voice and speech synthesizer voice.

IV. EXPERIMENTS AND RESULTS

A. An Example

Figure 5 gives an example of a dialog between a user and the system. *Aizuchi*, repetition, collaborative completion, and overlap are all included in the example. In Figure 5, first the system prompted a start-up utterance. Then, the user spoke “こんにちは (Hello)” and the system also spoke “はい こんにちは (Aha, Hello).” Next, the system spoke today’s weather to lead the user to the topic of weather. The system obtained the place where the user was (default value) and the weather around there, and kept the information in slots. With the next user utterance “最近雨ばかりだよな (Recently, it often rains, doesn’t it?)”, the system’s **collaborative completion** “多いですよね (rains a lot.)” was **overlapped**. The system detected the keywords/key phrases “最近 (recently)” and “雨 (rain)”, and knew that it had been raining. So, the system predicted that the user would say phrases that meant “it *often* rains” and tried to synchronize to the user with “多い (many)”. The system

has some response templates for collaborative completion and activates one of them if the user utterance and the current slot information meets a certain condition written as a decision rule. With the next user utterance “浜松の天気はどうですか (How about the weather in Hamamatsu?)”, the system detected a keyword “Hamamatsu (city name)” and responded immediately by the way of **repetition**. Then the system replied regarding the weather in Hamamatsu; “雨ばかりですよ (It always rains).” This dialog contained some chat-like dialog-specific phenomena such as *aizuchi*, repetition, and collaborative completion. Such phenomena often occur in human-to-human dialogs when the dialogs are warming up.

The results of the subjective evaluation are described in the next section. The prosody model is not evaluated during the subjective experiment with the dialog system. This is due to the fact that the prosody control by the model is not effective for a short dialog, as the model changes the prosodic information very slowly.

B. Evaluation of dialog system

1) *Setup*: The subjects used and evaluated the dialog system with the timing generator. Five different systems were evaluated, namely

System 1: the basic system (no overlap, *aizuchi* and repetition) using synthesized voices,

System 2: the basic system using recorded voices,

System 3: system embedded with the overlap function using recorded voices,

System 4: system incorporating all phenomena functions (that is, overlap, *aizuchi*, and repetition) using recorded voices, and

System 5: system incorporating all phenomena functions using synthesized voices.

We reveal whether or not there is a difference in the evaluation of timing according to a difference in voice quality. The subjects were instructed to concentrate on the evaluation of “timing” and “overlapping”. Using Systems 1 and 2, it compared the recorded voices with synthesized ones. For using System 3, it evaluated the overlap response. For using System 4, it evaluated *aizuchi* and repetition. For using System 5, it

TABLE II

RESULTS OF THE SURVEY QUESTION “WAS THE RESPONSE TIMING NATURAL OR UNNATURAL? (ON A SCALE OF 1 TO 5)” FOR EACH PHENOMENON. *Positive* SUBJECTS WERE THOSE WHO GAVE A 5 OR 4 POINT AS THEIR ANSWER, WHILE *negative* SUBJECTS WERE THOSE WHO GAVE A 1 OR 2 POINT AS THEIR ANSWER TO THE QUESTION. *Neutral* SUBJECTS WERE THOSE WHO GAVE A 3 POINT AS THEIR ANSWER TO THE QUESTION.

phenomenon	# of positive subjects	# of negative subjects	# of neutral subjects
ordinary response	8	1	1
overlap	4	3	3
aizuchi	6	4	0
repetition	2	6	2

TABLE III

RESULTS OF THE SURVEY QUESTION “IS IT EASY TO SPEAK IF THE SYSTEM INTRODUCES A PARTICULAR PHENOMENON? (ON A SCALE OF 1 TO 5)” FOR EACH PHENOMENON.

phenomenon	# of positive subjects	# of negative subjects	# of neutral subjects
overlap function	3	3	4
aizuchi	7	2	1
repetition	3	4	3

compared the recorded voices with synthesized ones, and also evaluated *aizuchi* and repetition.

Ten male subjects conversed on the dialog topic “weather information”, asking questions such as “Please get the weather information for various cities”. The subjects each had about 20 turns. After using a particular system, each subject completed a survey questionnaire, which included questions rated on a scale from one to five and open-ended questions.

2) *Questionnaire results*: Answers to the survey question “Was the response timing natural or unnatural?” are given in Table II for each phenomenon. This table indicates that the ordinary response obtained a very good evaluation, with *eight out of the ten users* convinced that the ordinary response was natural. Overlap and *aizuchi* also obtained a good evaluation. However, repetition timing was not considered by the subjects to be appropriate.

Answers to the survey question “Is it easy to speak if the system introduces a particular phenomenon?” for each phenomenon are given in Table III. This table indicates that the users were more comfortable with *aizuchi*.

With regards the effect of overlap in the questionnaire results, only three of the ten subjects answered that it was easy to speak with the introduction of the overlap function. They were of the opinion that “there is an overlap phenomenon in most natural conversation”. Conversely, three of the ten subjects felt that it was not easy to speak as “It is unpleasant when the system begins to speak while the user is still talking”. The overlap frequency of two of the three negative subjects was very high, 0.769 and 0.421 per utterance, respectively. It is thought that the impression of overlap was made worse because the system used overlap more than usual. It is thus necessary to adjust the frequency of overlap while talking, and to use it as a feature to decide the response. For the overlap frequency, the most frequent case is 0.769, and the least case is 0.045. The re-experiment was conducted to these two subjects, but the result did not change. The subject who was responded

by a few overlapping utterances from the system had very small intonation change in the utterance. Oppositely, the subject who was responded by a lot of overlapping utterances had a large intonation change in the utterance. The decision tree in our system uses prosodic information, so there is different in overlap frequency for the subjects.

With regards the effect of *aizuchi*, according to the results of the questionnaire, *seven of the ten subjects* answered that it was easy to speak after the system introduced *aizuchi*. These subjects were of the opinion that “It was friendly when there was *aizuchi*” and “It is understood that if there is *aizuchi*, the system is listening to the user’s utterance”. Conversely, there were two subjects who answered that it was not easy to speak. They concurred that “The timing of *aizuchi* was bad, so it was not easy to speak”.

With regards the effect of repetition in the results of the questionnaire, three of the ten subjects answered that it was easy to speak with the inclusion of the repetition function. They agreed that “It is good, because there is feedback from the system before the utterance of the phrase has been completed”. Conversely, four of the ten subjects answered that it was not easy to speak with the introduction of repetition. Opinions such as “It was not easy to speak, when the repetition was uttered during the user utterance” and “It is unpleasant when there is repetition with the recognition error” were quite contrary to the opinion of the subjects who were comfortable with the repetition. Some subjects liked the repetition and some did not, despite the occurrence of recognition errors. Since the opinion of the subjects is divided, it remains to be decided whether the system implements repetition according to user preferences.

As the effect of the voice quality (recorded v.s. synthesized) for the basic system, the number of subjects who answered as it was easy to speak for the recorded voices was *eight of ten subjects*. For the all phenomena functions implemented system, *nine of ten subjects* answered as it was easy to speak for the recorded voices. In each system, the subjects preferred the recorded voice system. *Eight of ten subjects* evaluated that the recorded voice was better than the synthesized voice for both (basic v.s. all phenomena functions implemented) systems. There were a lot of opinions “The recorded voice was more natural than the synthesized voice, and it feels so friendly”. According to [14], the synthesized voice was preferred in a simple system (in which the response timing is constant), and the recorded voice was preferred in a complex (such as all phenomena functions implemented) system. The balance is necessary for the voice quality and the quality of the conversation. However, it differed from the opinion, that is, the recorded voice system was preferred in both (basic, all phenomena functions implemented) systems in our experiment. We guess it was caused by the reason that the response timing of basic system is suitable.

With regards the effect of barge-in, *seven of the ten subjects* used barge-in and all of these had positive feedback. Opinions were varied: “Because the input can be corrected when a recognition error is found”, “It is possible to discover

TABLE IV
HIGH CORRELATION BETWEEN PHENOMENON AND SUBJECT'S
IMPRESSION

phenomenon	impression	correlation
overlap (positive)	recognition performance	0.765
overlap	easy to speak	-0.564
Aizuchi (positive)	correct response rate	0.648
Aizuchi	easy to speak	0.643

recognition errors early on”, and “The utterance can be spoken again soon”.

3) *Analysis of questionnaire results:* The correlation between the questionnaire results and dialog features (such as ASR performance, correct response rate, overlap frequency, and so on) was investigated. The high correlation features are given in Table IV.

Regarding the correlation between ASR performance (accuracy) and the questionnaire results, the overlap responses indicate a significant correlation (0.765, $p < 0.01^1$). The overlap response was thus preferred in the dialog with high ASR performance.

Regarding the correlation between the correct response rate and the questionnaire results, the *aizuchi* in the System 5 indicates a significance correlation (0.648, $p < 0.05$). Where a correct response is counted only to ordinary response, and it doesn't contain other responses (*aizuchi*, repetition, and collaborative completion). One of the authors judged whether the response was correct based on the conversation log or not. The correct response rate is defined as the rate of correct response to the user's requirement ($CorrectResponseRate = \frac{\#ofCorrectResponse}{\#ofOrdinaryResponse}$). The *aizuchi* was preferred in the dialog with high correct response rate. However, this trend was not indicated in the recorded human voices system (System 4). The difference of the correct response rate (System 2 - System 1) is correlated with the voice quality. The correlation coefficient was 0.647 ($p < 0.05$). The recorded human voices system was preferred by subjects in the basic system. This trend was not indicated in the all phenomena functions implemented system.

Regarding the correlation between the overlap frequency and the questionnaire results, the overlap indicates a high negative correlation (-0.564, $p = 0.090$) between actual overlap frequency and the subjective evaluation for overlap. The correlation coefficient has a large negative value, meaning that subjects preferred the system with low overlap frequency.

Regarding the correlation between *aizuchi* frequency and the questionnaire results, *aizuchi* in System 4 indicates a high significance correlation (0.643, $p < 0.05$) between actual *aizuchi* frequency and the subjective evaluation for *aizuchi*. In System 4, the system with high *aizuchi* frequency was preferred by the subjects.

V. CONCLUSIONS

In this paper, a dialog system utilizing real-time response generation and response timing generation was developed

¹The null hypothesis of no correlation was rejected at the significance level of 1%.

to perform chat-like friendly conversation. The phenomena occurring in human-to-human conversation (such as *aizuchi*, repetition, overlap, barge-in, etc.) were implemented in the system.

In the subjective evaluation, many subjects felt familiarity with *aizuchi*, and barge-in was also found to be useful. The naturalness of the timing generator was considered to be high for ordinary responses, overlap, and *aizuchi*. With respect to repetition, the subjects were divided into two conflicting groups. Thus, it remains to be decided whether the system implements repetition according to user preferences or not. With regards the overlap response, the user's impression deteriorated with a greater degree of overlap, but the overlap response was preferred in the dialog with high ASR performance.

In future work, the decision tree will be trained using dialogs between humans and the system. We also intend analyzing the effect of the conversations of two or more human agents with different personalities and characteristics.

REFERENCES

- [1] N. Ward and W. Tsukahara, “Prosodic features which cue back-channel responses in English and Japanese,” *Journal of Pragmatics*, 32, pp. 1177–1207, 2000.
- [2] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa, “Learning decision tree to determine turn-taking by spoken dialogue systems,” *ICSLP-02*, pp. 861–864, 2002.
- [3] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, “Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems,” *Journal of The Japanese Society for Artificial Intelligence*, vol. 20, no. 3, pp. 220–228, 2005.
- [4] R. Nishimura, N. Kitaoka, and S. Nakagawa, “Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling,” *Proceeding of the Interspeech 2008*, pp. 534–537, 2008.
- [5] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs,” *Language and Speech*, vol. 41, no. 3-4, pp. 291–317, 1998.
- [6] T. Ohsuga, M. Nishida, Y. Horiuchi, and A. Ichikawa, “Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue,” *Proceedings of Eurospeech2005*, pp. 33–36, 2005.
- [7] R. Nishimura, N. Kitaoka, and S. Nakagawa, “A spoken dialog system for chat-like conversations considering response timing,” *TSD 2007*, pp. 599–606, 2007.
- [8] K. Tanaka, S. Hayamizu, Y. Yamasita, K. Shikano, S. Itahashi, and R. Oka, “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.
- [9] R. J. Quinlan, “C4.5: programs for machine learning,” *Morgan Kaufmann*, 1992.
- [10] A. Kai and S. Nakagawa, “A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar,” *ICSLP-92*, pp. 257–260, 1992.
- [11] J. Zhang, L. Wang, and S. Nakagawa, “LVCSR based on context dependent syllable acoustic models,” *Asian Workshop on Speech Science and Technology, SP2007-200*, pp. 81–86, 2007.
- [12] J. Weizenbaum, “ELIZA — a computer program for the study of natural language communication between man and machine,” *Communications of the ACM* 9, no. 1, pp. 36–45, 1965.
- [13] S. Kawamoto, H. Shimodaira, T. Nitta, T. Nishimoto, S. Nakamura, K. Itou, S. Morishima, T. Yotsukura, A. Kai, A. Lee, Y. Yamashita, T. Kobayashi, K. Tokuda, K. Hirose, N. Minematsu, A. Yamada, Y. Den, T. Utsuro, and S. Sagayama, “Open-source software for developing anthropomorphic spoken dialog agent,” *Proc. of PRICAI-02, International Workshop on Lifelike Animated Agents*, pp. 64–69, 2002.
- [14] T. Itoh, N. Kitaoka, and R. Nishimura, “Subjective experiments on influence of response timing in speech dialogues,” *IPSI SIG Notes*, vol. 2008, no. 68, pp. 99–104, 2008. (in Japanese).