

A Spoken Dialog System for Spontaneous Conversations Considering Response Timing and Response Type

Ryota Nishimura^a, Non-member
Seiichi Nakagawa, Member

If a spoken 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 spoken dialog system that emulates human behavior in a dialog. The proposed system makes use of a decision tree to generate responses at the appropriate times. These responses include ‘*aizuchi*’ (back-channel), ‘repetition’, ‘collaborative completion’, etc. At each time interval, the decision tree generates the response timing features referring to the pitch and energy contours, recognition hypotheses, and the preparation status of the response generator. A subjective evaluation shows that there is a high degree of naturalness in the timing of ordinary responses and *aizuchi*, and that the spoken dialog system exhibits user-friendly behavior. The recorded voice system was preferred to a text-to-speech system (synthesized speech), and almost all subjects felt familiarity with the *aizuchi*. © 2010 Institute of Electrical Engineers of Japan. Published by John Wiley & Sons, Inc.

Keywords: spoken dialog system, speech communication, *aizuchi*, repetition, overlapping

Received 3 September 2009; Revised 21 June 2010

1. Introduction

Recently, there has been increased interest in and demand for interfaces providing automatic speech recognition (ASR). Because traditional systems provide no reaction to user utterances, a user cannot be certain that the system has recognized the utterance correctly. Considering this, several systems have been developed, such as the tutoring system [1] which incorporates back-channel and nodding, and the dialog robot [2] which deals with *aizuchi*, nodding, and facial expressions. Nevertheless, spoken dialog systems still tend to convey 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. According to Maynard [3], *aizuchi* are signals for the speaker to continue speaking, and also indicate the listener’s understanding of and agreement with the preceding utterances. Regarding *aizuchi* or turn-taking, the listener usually starts talking immediately after a short pause, or sometimes even overlaps the first speaker’s utterance. In natural human-to-human dialog, cooperative speech includes *aizuchi* and turn-taking with a variety of timings.

In smooth and cooperative human-to-human conversations, prosody information, including pitch and energy, is synchronized between the speakers. According to Kakita [4], if a speaker’s fundamental frequency (F0) is high in simple question and answer dialogs, the other speaker’s F0 also tends to be raised. Nagaoka *et al.* [5] showed that there is a positive correlation between switching pause durations for dialog partners. They suggest that enabling communication that is as smooth and natural as human-to-human dialogs is necessary to control the prosody of the system’s response appropriately. Previously, we analyzed human-to-human dialog to ascertain how prosody interacts between speakers, and then we

modeled the synchronizing tendency of F0, energy, and speech rate [6].

The purpose of this study is to generate natural responses, including *aizuchi*, repetition, collaborative completion, and ordinary responses, while also considering response timing. A decision tree that refers to prosodic information and surface linguistic information as features is employed to determine the appropriate response timing. Using this timing generation method, a human-friendly spoken dialog system has been developed [7]. The system discussed in this paper deals with all these phenomena, whereas the previous system described in the literature [7] cannot treat *repetition* and barge-in, and responds only after detecting a pause (in other words, the system cannot deal with overlapping responses). One of the aims of the proposed system is to make it appear so familiar to humans that they will want to chat with it.

This paper is organized as follows: Previous related works are described in Section 2. The novel architecture of the proposed spoken dialog system is presented in Section 3, while Section 4 describes the response-timing generation method and features required for the timing generation. An example of a dialog with the system is presented in Section 5. Section 6 provides experimental results of the spoken dialog system. Finally, we present our conclusions and suggest future works in Section 7.

2. Related Works on Spontaneous Conversation

Studies on *aizuchi* and turn-taking [8–11] indicate that pitch (F0) and energy are chiefly responsible for generating *aizuchi* and turn-taking. To date, various real-time *aizuchi* generation systems have been developed [12–14] 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 [15–17]. For example, Fujie *et al.* [18] used prosody information, especially F0 and the energy of the utterance, to determine appropriate timings for feedback responses. A finite state transducer-based speech recognizer was used to determine the sentence for feedback before the end of the utterance. While various kinds of responses need to be considered to emulate human

^a Correspondence to: Ryota Nishimura. E-mail: nishimura@slp.cs.tut.ac.jp
Department of Information and Computer Sciences, Toyohashi University of Technology, 1-1, Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan

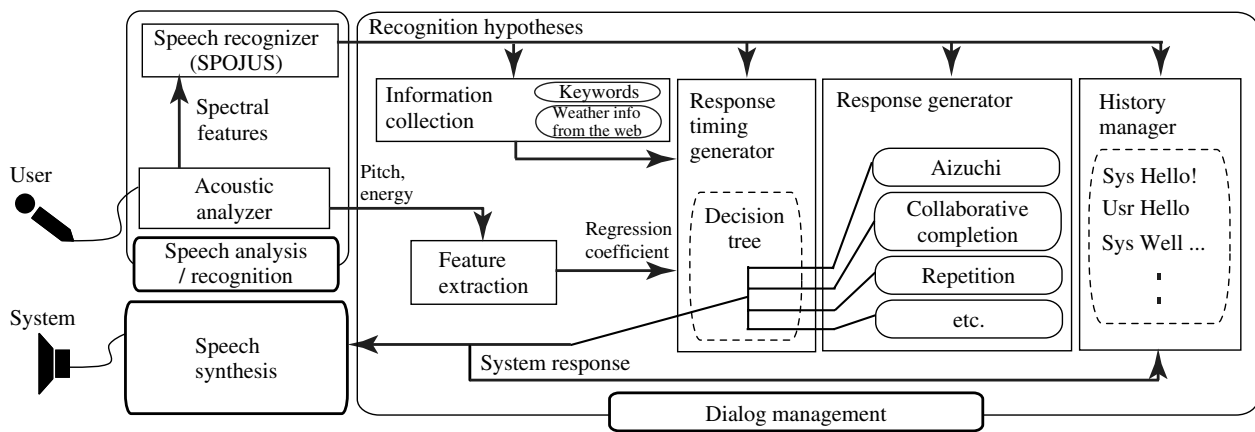


Fig. 1. Schematic diagram of the spoken dialog system

responses, these previous studies dealt only with a particular kind of response.

In this paper, we propose an integrated approach for the generation of various kinds of responses as follows:

- **Aizuchi:** In human-to-human conversations, at the completion of the speaker's turn, the listener either responds with content suggested by the speaker's utterance or starts the next turn thereby indicating that she/he has accepted what the first speaker said. To encourage the speaker to continue speaking in his/her turn, the listener responds with *aizuchi* to indicate that the speaker's utterances have been understood. In this case, explicit affirmative responses to questions/requests are not included in the *aizuchi*.
- **Repetition:** When a speaker senses that the listener cannot keep up with the conversation (e.g., when dictating a telephone number), the speaker divides the utterance into smaller fragments. In such cases, the listener often repeats the fragments to indicate an understanding thereof. Listeners also often repeat keywords appearing in the speaker's utterance. If a speaker wishes to confirm some of the contents of the utterance, that part of the utterance is given a rising tone to obtain the listener's agreement or correction, even before he/she has finished his/her turn.
- **Collaborative completion:** A listener often overlaps (synchronizes with) the speaker's utterances using the same content, or even complements the speaker's utterance by predicting the latter half of the utterance based on the first half so as to complete the utterance. This is called collaborative completion.
- **Ordinary response:** Various other response types (such as *greetings*, informed responses, etc.) are included in the ordinary response group.

Moreover, overlapping response timing is also treated.

- **Overlap:** In this paper, we refer to a system response that is uttered before the end of the user's previous utterance, as an 'overlap response'.
- **Barge-in:** Conversely, when a user utters a response before the end of the system's previous utterance, this is referred to as barge-in.

'Overlap' is a term that indicates the timing of the response from the system, whereas 'barge-in' is a term that indicates the timing of the input from the user.

Because our previous system [19] depended on pause detection to determine the timing for *aizuchi* and turn-taking, it could not deal with overlapping responses. The system proposed in this paper does not depend on pause detection and analyzes user utterances

continuously even while the user is speaking. This allows the system to deal not only with overlapping *aizuchi* and turn-taking (ordinary responses), but also with other types of responses, such as collaborative completion.

3. Spoken Dialog System

Figure 1 shows the novel architecture of the proposed spoken dialog system that can deal with the various phenomena described above. In this section, we give an overview of this system.

3.1. Speech analysis and recognition The speech recognizer SPOken Japanese Understanding System (SPOJUS) [20,21] 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. An example of a CFG is given below.

- SENTENCE: CITY DAY PRED
- CITY: city no
- no = {の (in)}
- city = {豊橋 (Toyohashi), 浜松 (Hamamatsu)}
- DAY: day no
- day = {今日 (today), 明日 (tomorrow)}
- PRED: pred
- pred: 天気はどうですか (How about the weather)
- pred: 天気は雨ですね (It is raining)

where CITY, DAY, and PRED are nonterminals, and no, city, day, and pred are terminal symbols.

SPOJUS uses 12 mel-frequency cepstrum coefficients (MFCCs), the first/second derivation of the MFCCs, and the first/second derivation of energy 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 and 10 ms, respectively. The Hidden Markov Model (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. The recognition speed is about $1.5 \times$ real time (RT) and 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. Moreover, at the same time, the system analyzes the input to extract prosodic information, such as pitch (F0) and energy, using a prosodic analyzer [18,22].

3.2. Dialog management Shown in Fig. 1 are details of the dialog manager, which is composed of five subcomponents ('information collection', 'feature extraction', 'response timing generator', 'response generator', and 'history manager') and generates response sentences using the hypotheses and prosodic information. One of the subcomponents, the 'response timing generator', uses a decision tree to determine the timing based on the features derived from the prosodic information. 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. Then, the system saves the information in 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.

3.2.1. Response generator The response generator prepares response sentences using an enzyme-linked immunosorbant assay (ELISA)-like procedure [23] with a slot-based history management, in addition to the recognition hypotheses. Thus, the response generator also serves as a simple dialog manager.

Our current system can handle *aizuchi*, *repetition*, *collaborative completion*, and various other *ordinary responses*. Thus, four patterns of response sentences are prepared simultaneously. Even while the user is speaking, the response generator continuously updates the response candidates using the intermediate hypotheses provided by the speech recognizer. *Aizuchi* can be used as a response for all input types, with the kind of *aizuchi* selected at random. Repetition is generated when a keyword (e.g., a city name in this system) is included in the input. For collaborative completion, for example, when the intermediate hypothesis is '最近あまり...' (Recently, it has not...), the system expects a user utterance taking into account recent weather information. And then, the system prepares the response '良くないね (That is not good)'. For ordinary responses, for example, when the intermediate hypothesis is '今日の (Today)', the system prepares the response '今日は暑いですね (It is hot today)'. However, if the intermediate hypothesis is '今日の豊橋 (in Toyohashi[city name], today)', the system changes the response to '今日の豊橋の天気は晴れです (Today it is fine in Toyohashi)'. Other simple sentences such as 'そうですね (right)' are also prepared and randomly selected as the response when no other appropriate sentences have been prepared by the templates. During a dialog, not only the keywords included in the user utterances, but also the current status of the weather retrieved from a web site (<http://www.imocwx.com/>) are kept in the slots and used for generating responses.

3.2.2. Response timing generator The response generator only constructs response sentences. To output a sentence, the response timing generator selects an appropriate sentence with the appropriate timing from the candidates prepared by the response generator. This timing generator also decides whether to respond or not and which response the system should make using a decision tree. Details of this process are discussed in Section 4.

3.3. Speech synthesizer To output responses as speech, we use either a recorded human voice or text-to-speech synthesized voice. A female voice was used to record 3410 sentences, including *greetings*, *aizuchi*, and weather information, with a familiar and lively intonation. GalateaTalk [24] is used as the speech synthesizer, and can control speaker type, voice tone (intonation), and speech rate. The same sentences have been recorded in both the human and synthesized voices.

4. Response Timing Generation

4.1. Features for timing generation According to Refs [8] and [11], the contour patterns of pitch and energy are related to the timing of response generation. For example, when pitch and/or energy 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 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). These are shown in Fig. 2. 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 shown in Ref. [11], the information for turn-taking is included in the last mora and in other parts of the utterance. Therefore, we used short and longer region (100 and 500 ms) from the last part of the utterance. As these coefficients can be calculated with very little computational cost, the calculation can be performed in real time.

Repetition and collaborative completion occur when a keyword of the conversation topic is input by the user [25]. When a speaker feels that the listener is unable to keep up with the conversation (e.g., when giving a telephone number), the speaker divides the utterance into several 'fragments'. In such cases, the listener often repeats the fragments to indicate his/her understanding. To imitate this behavior, the response generator should detect keywords in user utterances. In the recognition results (or intermediate hypothesis), an attribute is attached to each word, and this information is useful for detecting keywords. As explained in Section 3.2.1, 'keyword' here describes a word that is used to create the repetition or other response. Therefore, it is different from the usual meaning for 'keyword' that is the 'most important word in the conversation'. For example, 'How[*Question*] is[] the[] weather[*Weather*] in Hamamatsu[*city name*] today[*Date*]?''. Our present system focuses on weather information, and thus the attributes of keywords include place-names, dates, weather in topical places, etc. The attribute of the last word in the hypothesis (or intermediate hypothesis) is used as a feature.

The following features are used to implement the process above [26].

- Duration from the start of the user's preceding utterance
- Elapsed time from the end of the previous user utterance
- Elapsed time from the end of the previous system utterance
- Pitch/energy contour of the last 100 ms (consisting of three values)

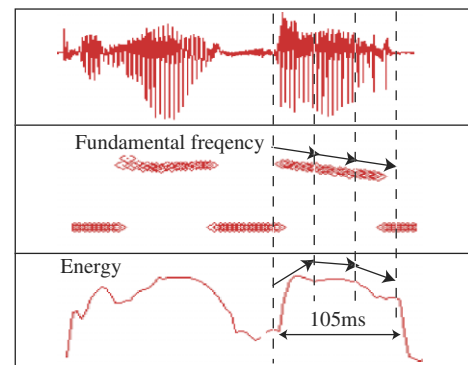


Fig. 2. Regression coefficients of fundamental frequency and energy at the end of an utterance

- Pitch/energy contour of the last 500 ms (consisting of five values)
- Attribute of the last word in the last recognition results (or current intermediate hypothesis).

The first item refers to ‘the duration of the user’s utterance plus pause length’ which implies the entire duration of the user’s utterance while it is being delivered. When the user’s utterance ends, the feature becomes ‘utterance duration plus pause length’. The latter feature refers only to the length of the pause which is 0, while the user is still delivering the utterance.

4.2. Response timing generation using a decision tree

Previously, we proposed a decision tree-based response timing generator [19], but this was only able to produce a response after detecting the pause (at the end of a 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. A part of the decision tree is shown in Fig. 3.

The response timing generator determines the response timing as well as selects the response sentence from the responses prepared by the response generator, using a decision tree based on the features introduced in Section 4.1. 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 the dialog act to be carried out by the system at every instance, from *aizuchi*, repetition, collaborative completion, ordinary response, and *wait*, as illustrated in Fig. 4. *Wait* means ‘do not output any response’. 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 many times as a response to the utterance. The state transition diagram for these dialog acts is depicted in Fig. 5. There are four types of responses, i.e. *aizuchi*, repetition, collaborative completion, and ordinary response. Collaborative completion and ordinary response are considered to be turn-taking. The priority of the response is determined by the decision tree using the prosodic information and does not change according to the content of the user input. It is beyond the scope of this paper to decide/change the priority of the response using the linguistic

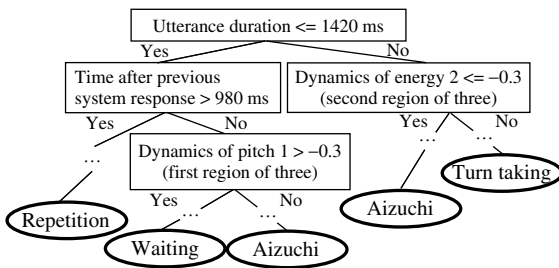


Fig. 3. Part of the decision tree

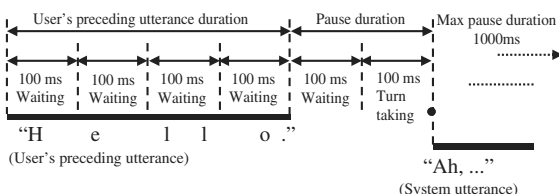


Fig. 4. Response timing generated using a decision tree

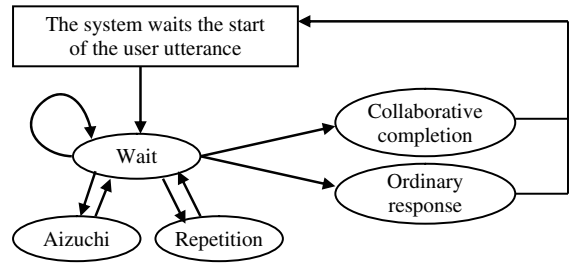


Fig. 5. State transition diagram for dialog acts

Table I. Frequency of phenomena in RWC corpus, where the value denotes the occurrence of each phenomenon per utterance

Phenomenon	Frequency
Overlap	0.190
Aizuchi	0.084
Repetition	0.003
Collaborative completion	0.004

information and this is left to a future work. The overlap indicates the timing of the response, so the existence of ‘repetition with overlap’, for example, is possible. The system decides whether it should respond at intervals of 100 ms. The dialog management does not have a dialog act for overlapping output, such as ‘response by overlap’. If the user is still uttering something when the system responds, this is called an overlap response. Therefore, ‘overlap’ is not included in Fig. 5.

The real world computing (RWC) corpus [27] was used to train the decision tree for *aizuchi*, turn-taking, and *wait*. It has 48 conversations each about 10-min long, giving a total of 6.5 h of dialog. The corpus consists of 16399 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 nonprofessional men and women. C4.5 [28] was used to construct the decision tree.

On the basis of the decision tree, *aizuchi* occurs when 2 or more seconds have elapsed since the latest response of the system, when 0.5 or more seconds have elapsed since the start of the user utterance, and when the pitch contour is flat or has a negative slope. With regards the other phenomena, namely repetition and collaborative completion, there were insufficient training data in the corpus. When training takes place for turn-taking ‘overlap’ is not explicitly provided for. ‘Elapsed time from the end of the previous user utterance (pause length)’ in the feature discussed in Section 4.1 is set to 0 ms. This feature is, in fact, set to 0 ms, while the user is in the process of delivering an utterance. The frequency of each phenomenon is given in Table I. Consequently, we manually added some rules. On the basis of the hand-crafted rules, repetition occurs when 2 or more seconds have elapsed since the latest response by the system, and when the last word in the recognition hypothesis is a city name. Through various conversations with the system, we determined the decision rules heuristically. Consequently, these rules may be dependent on the domain. Collaborative completion occurs immediately after the response has been prepared (i.e., the user input matches the template and the system response has been prepared).

In addition, the system relies on an exceptional rule to continue the dialog, in that the system prompts the user to say something after a long pause (6 s in our system). Furthermore, when a pause of over 1000 ms occurs since the last user utterance, the system gives an arbitrary response to the user without consulting the decision tree.



Fig. 6. Example dialog between the system and a user

5. Domain and Example Dialog with the System

5.1. Domain Weather information was chosen as the dialog domain for the system by the following three reasons. Firstly, many subjects can talk comfortably about this domain, and secondly, it incorporates real information from the WEB. The final reason is that the domain and utterances (or grammar and vocabulary) can be restricted naturally.

5.2. Example An example dialog with the system is illustrated in Fig. 6. The top part of the figure shows the user utterances, and the bottom the system responses.

According to Fig. 6, first the system prompts with a start-up utterance. Then, the user utters ‘こんにちは (Hello)’ and the system replies ‘はい こんにちは (Aha, Hello)’. Next, to lead the user to the topic of weather, the system gives today’s weather, having first ascertained the location of the user (default value) and the weather in that area, and placed this 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 (rains a lot)’ overlaps. By detecting the keywords/key phrases ‘最近 (recently)’ and ‘雨 (rain)’, the system knows that it has been raining. Thus, the system predicts that the user will say something along the lines of ‘it often rains’ and tries to synchronize with the user with ‘多い (a lot)’. The system has some response templates for collaborative completion and activates one of these if the user utterance and the current slot information meet 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 detects a keyword ‘Hamamatsu (city name)’ and responds immediately with **repetition**. Thereafter, the system replies regarding the weather in Hamamatsu; ‘雨ばかりですよ (It always rains)’. This dialog contains some dialog-specific phenomena such as *aizuchi*, repetition, and collaborative completion. Such phenomena often occur in human-to-human dialog when the dialog is starting up.

As described above, the proposed system can work with many kinds of phenomena appearing in natural human-to-human spoken dialog including overlapping utterances when given appropriate rules, templates, and parameters.

6. Experiments and Results

6.1. Speech recognition performance The recognition accuracy is defined by the following expressions:

$$\text{Cor} = \frac{\text{Len} - \text{Sub} - \text{Del}}{\text{Len}} \times 100[\%] \quad (1)$$

$$\text{Acc} = \frac{\text{Len} - \text{Sub} - \text{Del} - \text{Ins}}{\text{Len}} \times 100[\%] \quad (2)$$

where Cor is the correct ASR, Acc is the ASR accuracy. Len is the number of words in the target string (utterance), Sub is the number of substitutions compared with the target string, Del is the

Table II. ASR performance for dialog (spontaneous) speech in a weather information task

Speaker	Cor (%)	Acc (%)	OOV (%)
Speaker 1	52.2	47.7	11.2
Speaker 2	43.4	41.5	4.0
Speaker 3	69.1	66.6	1.7
Speaker 4	69.7	66.0	0.9
Speaker 5	77.3	75.4	0.0
Speaker 6	48.8	45.6	0.4
Speaker 7	42.2	35.8	10.1
Speaker 8	28.7	25.9	22.7
Speaker 9	56.1	40.4	0.4
Speaker 10	79.7	76.5	0.0
Average	56.7	52.1	5.1

number of deletions compared with the target string, and Ins is the number of insertions compared with the target string.

The ASR performance of our speech recognizer for dialog (spontaneous) speech in the weather information task is given in Table II. We used ten male subjects in their twenties, all enrolled for Bachelor or Master degrees. This domain has a vocabulary of 158 (the number of city names was restricted and the vocabulary reduced). It is very difficult for the ASR system to recognize spontaneous speech in a real environment, because spontaneous speech is disfluent and ambiguous, and the environment can be noisy. Moreover, in natural and smooth dialog, the disfluency and ambiguity become more pronounced. The average recognition accuracy in all dialogs is 52.2%, while the average for each speaker is between 76.5 and 25.9%. The out of vocabulary (OOV) rate is shown in the rightmost column of Table II. From these results, it is evident that speakers with bad recognition rates (Speakers 7 and 8) have high OOV rates. Worse results were obtained when the subjects uttered phrases that did not conform to the CFG or were out of the domain.

6.2. Evaluation of response selection We evaluated the responses selected by the decision tree using the RWC corpus. Acts evaluated are *aizuchi*, turn-taking, and *wait*. As mentioned in Section 4.2, there were insufficient training data for repetition and collaborative completion in the corpus, with the result that these phenomena were not evaluated. The decision tree is evaluated using a fourfold cross validation, according to the four sub-corpora in the RWC corpus (namely, male, female; ‘car sales’, ‘overseas trip planning’). All the utterances in a corpus are used for the system model’s learning. Four kinds of corpora, namely, ‘car sales (male)’, ‘car sales (female)’, ‘overseas trip (male)’, and ‘overseas trip (female)’ are employed. The system model takes into consideration the utterances of both the salesperson and customer in a single corpus, and the model is then evaluated using all the utterances of the other three corpora. Therefore, the system

Table III. Evaluation of response selection (at every segment, data open test)

Measure	Aizuchi (%)	Turn-taking (%)	Wait (%)
Recall	53.2	58.9	98.9
Precision	49.5	44.8	99.5
F-measure	51.3	50.9	99.2

model is independent of the corpus and is learned using several persons' utterances. The evaluations do not include speaker-open conditions, as the car sales person or the trip planner is the same person in the particular corpus used. Unfortunately, no other useful corpora exist for speaker-open conditions. The experiment for domain-open condition was not conducted because corpora of a specific domain should be collected when the system is actually constructed. The answer response is determined at every segment (period of 100 ms) as shown in Fig. 4. The response from almost all segments was determined to be *Wait*, and only those segments with the response timing (not *Wait*) produced an answer such as *aizuchi* and so on. The results are summarized in Table III.

In Table III, for the actual response in the corpus, the time from the start of an utterance until the response by the other party (i.e., the duration of the utterance plus pause) was divided into 100-ms segments, and each segment was judged. A similar evaluation method is used in Ref. [7], but in this previous work it was used only to judge the pause.

According to Table III, the F-measure for *aizuchi* is 51.3%. This appears to be an inferior result in a reproduction experiment. It should be noted, however, that these results do not imply that our method is defective. Humans also differ in their individual response (*aizuchi*) generation. As pointed out in Ref. [7], a different person to the one in the corpus is able to recall only 50% of the *aizuchi* in the corpus when hearing one side of the dialog speech. These rates are comparable with the rates in Table III. This indicates that the rate of agreement between human speakers is not so high, and that the low agreement does not imply any lack of naturalness. Therefore, we cannot evaluate the generator using only this objective test, and the timing of our generator may in fact be natural. We will confirm this in the following sections.

6.3. Evaluation of timing generation The naturalness of the timing generated by the system was evaluated subjectively. Here, only *aizuchi* and turn-taking were evaluated, because repetition and collaborative completion were the phenomena with few occurrences.

To evaluate the timing of the generator, we prepared samples of *aizuchi* and turn-taking with timing generated by the decision tree.

We inserted an *aizuchi* extracted from a part of a dialog at the *aizuchi* timing point generated by our timing generator. We also created samples of turn-taking. Thus, we chose some filled pauses, such as 'Ettodesune' ('Well... let's see' in English), to insert at the time of the system output. The filled pause is used to create a sample for evaluating turn-taking. Subjects listened to the inserted *aizuchi* with one preceding sentence and evaluated only the timing.

We compared the timing by the generator with that in the corpus. In actual dialogs from the corpus, the responses may have a certain meaning consistent with the dialog context and this may appear more natural to the subjects, especially in the case of turn-taking. To ensure that the subjects evaluated only the timing, we also replaced the *aizuchi* or filled pauses in the actual responses with *aizuchi* or a filled pause extracted from other parts of the dialog, as in the case of the generator [7,29]. The information (various prosodic features) included in human speech is lost in synthesized speech. Therefore, some prosodic information (such as pause length and pitch/energy contour) used by the system cannot be evaluated. Consequently, we used human (real-dialog) speech for the evaluation. We created 20 samples for each phenomenon. The five subjects listened to these sample voices and completed the relevant questionnaires, using the following rating scale (1: too early; 2: early; 3: good; 4: late; 5: too late; and 0: outlier).

The results are shown in Table IV. The 'naturalness' in the table indicates the rate of 'goodness'. In the table, the naturalness of the decision tree timing is comparable to the naturalness of the corpus (i.e., human-to-human dialog) timing. Here, the balance of early and late responses was different from the decision tree and the corpus. Determining the timing of the decision tree output by subjects is influenced by the content of the utterance. In the decision tree used in this study, the content of the utterance is not considered. Our decision tree is learned in such a way that the optimal rate for the decision tree is the same as the rate used in the corpus. Therefore, the balance does not agree with that in the corpus. The goal of machine learning is to obtain 100% 'good' decisions. We assumed that the timing of the corpus was correct, although it might be incorrect for the subjective evaluation. In other words, the target was not learned to reproduce the results of the evaluation of the subjectivity of the corpus. The subjective evaluation was thus influenced by the content of the utterance.

6.4. Evaluation of the spoken dialog system

6.4.1. setup The subjects used and evaluated the spoken dialog system with the timing generator. Five different systems, as described below, were evaluated.

System 1: the basic system (no overlap, *aizuchi* or 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 (i.e., overlap, *aizuchi*, and repetition) using recorded voices;

System 5: system incorporating all phenomena 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 'overlap' only.

Ten male subjects in their twenties participated in the experiment, and the subjects enrolled for Bachelor or Master degrees. (These were the same subjects used in the experiment documented in Table II.) The subjects conversed on the topic 'weather information', asking questions such as 'Please get the weather information

Table IV. Evaluation of response timing by subjective evaluation

Phenomenon	Timing	Too early	Early	Good	Late	Too late	Outlier	Naturalness (%)
Aizuchi	Decision tree	0	6	61	20	11	2	61.0
	Corpus	14	26	58	2	0	0	58.0
Turn-taking	Decision tree	9	26	53	9	2	1	53.0
	Corpus	7	31	51	10	0	1	51.0

for various cities'. Each subject had about 20 turns with each system. After using a particular system, the subject completed a survey questionnaire, which included questions rated on a scale from one to five and open-ended questions. The order in which the spoken dialog systems were used was fixed for the ten subjects, i.e. sequentially from Systems 1 to 5. The reason for fixing the order was to reduce any influence caused by an order change. If the order were not fixed, the number of combinations would increase making the experiment more complicated. Moreover, if users repeatedly use spoken dialog systems, they become accustomed to the system in accordance with their usage thereof, which would mean that we could not evaluate the spoken dialog system fairly. Therefore, when using System 5 users are mostly accustomed to it. The impression of the current system is affected by the impression of the previous system encountered by the user. The order is also important, as the evaluation of a system may be biased by different orders. We carefully designed the order so that it would not affect the system. Furthermore, since the number of experimental trials for each subject was small, the experiential (or habitual) effect was also small.

The questionnaires used are given in the Appendix. For Systems 1 and 2, we compared the recorded voices with the synthesized ones. For System 3, we evaluated the overlap response. For System 4, we evaluated *aizuchi* and repetition. For System 5, we compared the recorded voices with the synthesized ones, and also evaluated *aizuchi* and repetition.

6.4.2. Questionnaire results Answers to the survey questions are given in Table V. *positive* subjects are those who gave a 5- or 4-point rating as their answer, while *negative* subjects are those who gave a 1- or 2-point rating as their answer to the question. *Neutral* subjects are those who rated the question as 3. With regards Q1-2(2), Q2-3(2), Q3-3(2), and Q4-3(2), *positive* indicates a fast response timing, while *negative* indicates a slow response timing. The timing evaluation (Q2-3(1)) of the overlap response is an evaluation of the timing of the overlap only, while the evaluation (Q1-2(1)) of the baseline is an evaluation of the entire timing. So it is not possible to compare these directly. The evaluation for Q2-3(1), Q3-7(1), and Q4-7(1) is 3.0 or more, confirming that the proposed timing mechanism works well.

Regarding the effect of overlap in the questionnaire results, only four of the ten subjects agreed that it was easy to communicate with the introduction of the overlap mechanism (Q2-4). 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.

With regard the effect of *aizuchi*, according to the results of the questionnaire, six of the ten subjects agreed that it was easy to communicate after the system introduced *aizuchi* (Q3-4 for *aizuchi*). These subjects were of the opinion that 'It was more friendly with *aizuchi*' and 'It is generally understood that if there is *aizuchi*, the system is listening to the user's utterance'. Conversely, there were three subjects who answered that it was not easy to speak. They concurred that 'The timing of *aizuchi* was bad, making it difficult to speak'.

Regarding the effect of repetition in the results of the questionnaire, three of the ten subjects answered that it was easy to communicate with the inclusion of the repetition function (Q3-4 for repetition). 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

Table V. Average ratings from the questionnaire, and the number of positive, negative, and neutral subjects out of the ten

Questionnaire	Average rating	Number of positive subjects	Number of negative subjects	Number of neutral subjects
Q1-1	4.3	8	(O)1	1
Human voice				
Q1-2(1)	3.8	7	2	1
Q1-2(2)	2.9	2	(A)3	5
Synthesized voice				
Q1-2(1)	3.4	5	3	2
Q1-2(2)	2.8	1	(A)3	6
Q2-1	3.6	7	2	1
Q2-2	2.5	1	5	4
Q2-3(1)	3.2	4	(A, O)3	3
Q2-3(2)	3.5	4	0	6
Q2-4	3.2	4	(A, O)3	3
For <i>aizuchi</i>				
Q3-1	4.3	8	(A)1	1
Q3-2	3.0	2	(A)2	6
Q3-3(1)	3.1	6	(A, O)4	0
Q3-3(2)	3.3	3	1	6
Q3-4	3.4	6	(A, O)3	1
For repetition				
Q3-1	4.2	8	(A)1	1
Q3-2	2.7	2	(A, O)3	5
Q3-3(1)	2.5	2	(A, O)6	2
Q3-3(2)	3.5	4	1	5
Q3-4	3.0	3	(A, O)4	3
Q3-7(1)	3.6	8	1	1
Q3-7(2)	3.1	2	(A, O)2	6
For <i>aizuchi</i>				
Q4-1	4.0	9	(A)1	0
Q4-2	2.8	1	(A)3	6
Q4-3(1)	3.4	6	(A)2	2
Q4-3(2)	2.8	2	(A)2	6
Q4-4	3.8	7	(O)2	1
For repetition				
Q4-1	3.7	7	2	1
Q4-2	2.8	3	(A)4	3
Q4-3(1)	2.6	1	(A, O)5	4
Q4-3(2)	2.9	4	(A)4	2
Q4-4	2.9	3	(A, O)4	3
Q4-7(1)	3.2	5	(A, O)4	1
Q4-7(2)	3.2	4	(A)2	4
Q4-8	4.5	9	1	0

Notes: Two markers (A, O) indicate whether those subjects with the worst ASR/overlap performance, respectively, are included in the negative subjects. The 'A' depicts the two subjects with the worst ASR performance (Speakers 7 and 8 in Table II), while the 'O' depicts the subject to whom the system responded with a lot of overlapping utterances (overlap frequency of 0.769).

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 a 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 was divided, it remains to be decided whether or not the system implements repetition according to user preference.

As for the effect of voice quality (recorded vs. synthesized) for the basic system (Q1-1), eight out of the ten subjects answered that it was easy to communicate with the recorded voices. For the system implementing all phenomena functions, nine of the ten subjects answered that it was easy to communicate with the

recorded voices (Q4-8). In each system, the subjects preferred the recorded voice system. Eight out of the ten subjects agreed that the recorded voice was better than the synthesized voice in both (basic, as well as fully implemented phenomena) systems. There were many varied opinions: ‘The recorded voice was more natural than the synthesized voice, and it feels so friendly’. According to Ref. [30], the synthesized voice was preferred in a simple system (in which the response timing is constant), whereas the recorded voice was preferred in a complex system (such as that implementing all phenomena functions). The balance is necessary for the voice quality and the quality of the conversation. The results in our experiment, however, differed in that the recorded voice was preferred in both systems (basic and fully implemented phenomena). The reason for this is that the response timing of the basic system is adequate.

Regarding the effect of barge-in, seven of the ten subjects used barge-in and all of these had positive feedback. Opinions for the reasons were varied: ‘The input can be corrected when a recognition error occurs’, ‘It is possible to discover recognition errors early on’, and ‘The utterance can be spoken again soon’. The subjects can immediately correct recognition errors in the system, and can thus raise the efficiency of the conversation by implementing the barge-in.

The overlap frequency (System 3) and *aizuchi* and repetition frequencies (Systems 4 and 5) per utterance are given in Table VI. Regarding the overlap frequency, the frequency for the subject with the most overlap is 0.769, and with the least is 0.045. The experiment was redone for these two subjects, but without any change in the results. The subject with the least overlapping responses from the system exhibited very little intonation change in his/her utterances. Conversely, the subject with the most overlapping responses exhibited large intonation changes in his/her utterances. The decision tree in our system uses prosodic information, so there is a difference in overlap frequencies for the subjects. Comparing Tables I and VI, the frequency of the output by the decision tree is higher than that by the corpus. In Table VI, the repetition frequency, in particular, increases greatly. This is due to the fact that the system often repeats the city name based on the rules. The value is different between the corpus and the decision tree. However, the result of the evaluation of the decision tree output is good, thereby confirming that the decision tree works well.

The results from the four subjects with the highest recognition accuracy are summarized in Table VII. Their ASR recognition accuracies were higher than 60%, and the average recognition accuracy was 71.1%. The average ratings for the questions and the number of positive and negative subjects are given in the table. The average ratings for questions Q2-3(1), Q3-7(1), and Q4-7(1) are 4.0 or more. Therefore, it is clear that if the ASR goes well, the proposed timing mechanism also works well. The four subjects included in Table VII have the top four ASR results, as well as the top four correct response rates (CRR). In the experiment, these subjects carried out the task in the ideal environment we assumed. Therefore, despite there only being four subjects, these findings are useful.

The results of the questionnaire for the overall evaluation of the system (Q5-1 and Q5-3) are given in Table VIII. System 4 (implementing all functions and using recorded human voices)

Table VI. Frequency of response type

Phenomenon	Average	Min	Max
Overlap	0.240	0.045	0.769
Aizuchi	0.255	0.087	0.393
Repetition	0.100	0.045	0.182

Note: Values denote the occurrence of each phenomenon per utterance.

Table VII. Average ratings from the questionnaire and the number of positive, negative, and neutral subjects

Questionnaire	Average rating	Number of positive subjects	Number of negative subjects	Number of neutral subjects
Human voice				
Q1-2(1)	4.5	4	0	0
Synthesized voice				
Q1-2(1)	3.5	2	1	1
Q2-3(1)	4.3	3	0	1
Q3-7(1)	4.0	4	0	0
Q4-7(1)	4.0	3	0	1

Notes: Only the four subjects with the highest recognition accuracy (greater than 60%) are included. The average recognition accuracy is 71.1%.

Table VIII. Results of Questionnaire 5 (Q5-1 and Q5-3)

Question	System ID				
	1	2	3	4	5
Q5-1	0	2	2	5	1
Q5-3	4	1	2	0	3

made a good impression on the subjects. There was a low evaluation for System 1 which is a simple system. System 5 was also given a low evaluation. It appears that the evaluation of a system tends to decrease with poor voice quality, even in the case of a high-performance system.

6.4.3. Analysis of questionnaire results We investigated the correlation between the questionnaire results and dialog features (such as ASR performance, correct response rate, overlap frequency, and so on).

Regarding the correlation between ASR performance (Acc) and the questionnaire results, the overlap responses (Q2-4) indicate a significant correlation (0.765, $p < 0.01^1$). The overlap response was thus preferred in dialogs with high ASR performance.

Regarding the correlation between the correct response rate and the questionnaire results, the *aizuchi* (Q4-4 for *aizuchi*) in System 5 indicates a significant correlation (0.648, $p < 0.05$). Here, a correct response is counted only as an ordinary response, and does not contain other response types (*aizuchi*, repetition, or collaborative completion). One of the authors decided whether or not the response was correct based on the conversation log. The correct response rate is defined as the rate of correct responses according to the user’s desires (correct response rate = $\frac{\text{number of correct responses}}{\text{number of ordinary responses}}$). *Aizuchi* was preferred in dialogs with a high correct response rate. However, this trend was not observed in the recorded human voice system (System 4).

Regarding the correlation between the overlap frequency and the questionnaire results, the overlap (Q2-4) indicates a high 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* (Q3-4 for *aizuchi*) in System 4 indicates a high significant correlation (0.643, $p < 0.05$) between actual *aizuchi* frequency and the subjective evaluation for *aizuchi*. In the recorded human voice system (System 4), the system with high *aizuchi* frequency was preferred by the subjects. However,

¹ The null hypothesis of zero correlation was rejected at the significance level of 1%.

this trend was not observed in the synthesized voice system (System 5).

As mentioned above, the ASR performance, CRR, overlap frequency, and aizuchi frequency are related to the results of the subjective evaluations. Subjective evaluations with a high ASR performance also show a high overlap and aizuchi frequency for Q2-3(1), Q2-4, Q3-1, Q3-3(1), and Q3-4. The factors mentioned previously affect the subjective evaluation, but the actual effect thereon differs depending on the ASR performance.

7. Conclusions

In this paper, a spoken dialog system utilizing real-time response generation and response timing generation was developed to engage in friendly conversation. Phenomena occurring in human-to-human conversation, such as *aizuchi*, repetition, collaborative completion, overlap of response timing, and barge-in by the user, among others, were implemented in the system.

The naturalness of the decision tree-based timing generator is comparable with human conversation while the behavior of the spoken dialog system gives a user-friendly impression. In the subjective evaluation, the recorded voice system was preferred to the synthesized voice system, and many subjects felt familiarity with *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 the user's preferences. Collaborative completion is a phenomenon that often occurs in human-to-human conversations. However, collaborative completions do not occur if the conversation does not flow. In the framework of the dialog system dealt with in this paper, collaborative completions can be achieved, but it is difficult for the dialog system to put together a collaborative completion in an actual conversation, because the system needs to predict the linguistic information in order to output the collaborative completion. Moreover, it is necessary to prepare a response corresponding to the predicted information to create the collaborative completion. The proposed system is not able to predict linguistic information, and the response templates are limited to the range of conversation expected by the developer. Although a collaborative completion can be output by this system framework, the analysis thereof has not been done because it does not occur in the subjective evaluations.

Regarding the overlap response, user impressions deteriorated with a greater degree of overlap. The overlap response was preferred in dialogs with high ASR performance. With good ASR performance, the proposed timing mechanism works well, and the users were impressed by the phenomenon. Subjects however, preferred a system with low overlap frequency. *Aizuchi* are preferred in dialog with a higher correct response rate in the synthesized voice system, whereas more frequent *aizuchi* are preferred in the recorded voice system. With the implementation of the barge-in, users can immediately correct any recognition errors by the system, thereby improving the quality of the conversation.

Various phenomena considered in this paper are observed in human-to-human conversation regardless of the domain. Therefore, neither the system framework proposed in this paper, nor the evaluation result depends on the domain. The application of the tutorial system and conversation robot mentioned in Section 1 is also possible, although the dialog management part would become complex. Although the frequency and timing of the conversational phenomena may differ from domain to domain this system framework can be ported to another domain by training the decision tree using an appropriate corpus for the application domain.

In future work, we intend to continue using dialogs between humans and the system to train the decision tree. We also aim to adopt prosodic synchrony to make the system response more natural in longer dialogs with many turns (for shorter dialogs, this has

almost no effect). To apply this system to a more complex environment and an actual experiment will be a focus of our future work. We also intend to analyze conversations between two or more humans/agents with different personalities and characteristics.

Appendix: Questionnaires

- Questionnaire 1 (for Systems 1 and 2)
 - Q1-1** Which system is easier to communicate with us?
 - Synthesized voices (1 2 3 4 5) Human voices
 - Q1-2** response timing
 - (1) unnatural (1 2 3 4 5) natural
 - (2) slow (1 2 3 4 5) fast
 - Q1-3** impression
 - open question
- Questionnaire 2 (for System 3)
 - Q2-1** Do you feel the overlap response?
 - doesn't feel (1 2 3 4 5) feels a lot
 - Q2-2** frequency of overlap response
 - too few (1 2 3 4 5) too many
 - Q2-3** overlap response timing
 - (1) unnatural (1 2 3 4 5) natural
 - (2) slow (1 2 3 4 5) fast
 - Q2-4** it is easy to speak with the introduction of overlap response
 - not easy (1 2 3 4 5) easy
 - Q2-5** Why did you select the rating in the previous question?
 - open question
 - Q2-6** impression
 - open question
- Questionnaire 3 (for System 4)
 - Similar questions as in [Questionnaire 2], with overlap replaced first by *aizuchi* and then by repetition. One additional question as given below.
 - Q3-7** timing of ordinary responses
 - (1) unnatural (1 2 3 4 5) natural
 - (2) slow (1 2 3 4 5) fast
- Questionnaire 4 (for System 5)
 - The same questions as in [Questionnaire 3]
 - One additional question as given below.
 - Q4-8** Which system is easier to communicate with us?
 - Synthesized voices (1 2 3 4 5) Human voices
- Questionnaire 5 (for all systems)
 - Q5-1** With which system is it the easiest to conduct a dialog?
 - (System 1–System 5)
 - Q5-2** Why did you select the particular answer in the question above?
 - open question
 - Q5-3** Which system is not the easiest to communicate with us?
 - (System 1–System 5)

Q5-4 Why did you select the particular answer in the question above?

- open question

Q5-5 impression

- open question

References

- (1) Rajan S, Craig S, Gholson B, Person N, Graesser A, TRG. Autotutor: incorporating back-channel feedback and other human-like conversational behaviors into an intelligent tutoring system. *International Journal of Speech Technology* 2001; **4**:117–126.
- (2) Fujie S, Kobayashi T. Realization of rhythmic dialogue on spoken dialogue system using para-linguistic information, *4th Joint Meeting Acoustical Society of America and Acoustical Society of Japan*, Hawaii, 2006, 534–537.
- (3) Maynard SK. *Kaiwa Bunseki*. Kuroshio Shuppan: Tokyo, Japan, 1993; (in Japanese).
- (4) Kakita K. Inter-speaker interaction of F0 in dialogs, *Proceedings of ICSLP-1996*, 1996, 689–692.
- (5) Nagaoka C, Komori M, Yoshikawa S. Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction, *Proceedings of Active Media Technology 2005 (AMT-2005)*, 2005, 529–534.
- (6) Nishimura R, Kitaoka N, Nakagawa S. Analysis of relationship between impression of human-to-human conversations and prosodic change and its modeling, *Proceeding of the Interspeech 2008*, 2008, 534–537.
- (7) Kitaoka N, Takeuchi M, Nishimura R, Nakagawa S. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Journal of the Japanese Society for Artificial Intelligence* 2005; **20**(3):220–228.
- (8) Koiso H, Horiuchi Y, Tutiya S, Ichikawa A, Den Y. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech* 1998; **41**(3–4):291–317.
- (9) Gelykens R, Swerts M. Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication* 1994; **15**:69–77.
- (10) Hirschberg J. Communication and prosody: functional aspects of prosody. *Speech Communication* 2002; **36**:31–43.
- (11) Ohsuga T, Nishida M, Horiuchi Y, Ichikawa A. Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue, *Proceedings of Eurospeech 2005*, 2005, 33–36.
- (12) Ward N, Tsukahara W. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 2000; **32**:1177–1207.
- (13) Okato Y, Kato K, Yamamoto M, Itahashi S. Insertion of interjectory response based on prosodic information, *IEEE Workshop Interactive Voice Technology for Telecommunication Applications (IVTTA-96)*, 1996, 85–88.
- (14) Noguchi H, Den Y. Prosody-based detection of the context of backchannel responses, *Proceedings of ICSLP-98*, 1998, 487–490.
- (15) Sato R, Higashinaka R, Tamoto M, Nakano M, Aikawa K. Learning decision tree to determine turn-taking by spoken dialogue systems, *ICSLP-02*, 2002, 861–864.
- (16) Hirasawa J, Nakano M, Kawabata T, Aikawa K. Effects of system barge-in responses on user impressions. *EUROSPEECH-99*, 1999; **3**:1391–1394.
- (17) Kamm C, Narayanan S, Dutton D, Ritenour R. Evaluating spoken dialogue systems for telecommunication services, *Eurospeech-97*, 1997, 2203–2206.
- (18) Fujie S, Fukushima K, Kobayashi T. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system, *Interspeech-05*, 2005, 889–892.
- (19) Takeuchi M, Kitaoka N, Nakagawa S. Timing detection for realtime dialog systems using prosodic and linguistic information, *Speech Prosody 2004*, 2004, 529–532.
- (20) Kai A, Nakagawa S. A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar, *ICSLP-92*, 1992, 257–260.
- (21) Zhang J, Wang L, Nakagawa S. LVCSR based on context dependent syllable acoustic models, *Asian Workshop on Speech Science and Technology, SP2007-200*, 2007, 81–86.
- (22) Goto M, Ito K, Hayamizu S. A real-time filled pause detection system for spontaneous speech recognition, *Eurospeech-99*, 1999, 227–230.
- (23) Weizenbaum. J. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 1965; **9**(1):36–45.
- (24) Kawamoto S, Shimodaira H, Nitta T, Nishimoto T, Nakamura S, Ito 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, Sagayama S. Open-source software for developing anthropomorphic spoken dialog agent, *Proceedings of PRICAI-02, International Workshop on Lifelike Animated Agents*, 2002, 64–69.
- (25) Ishizaki M, Den Y. *Danwa to Taiwa*. Tokyo Daigaku Shuppankai: Tokyo, Japan, 2001 (in Japanese).
- (26) Nishimura R, Kitaoka N, Nakagawa S. A spoken dialog system for chat-like conversations considering response timing, *TSD 2007*, 2007, 599–606.
- (27) Tanaka K, Hayamizu S, Yamasita Y, Shikano K, Itahashi S, Oka R. Design and data collection for a spoken dialogue database in the real world computing program, *Proceedings of ASA-ASJ Third Joint Meeting*, 1996, 1027–1030.
- (28) Quinlan RJ. *C4.5: Programs for Machine Learning*. Morgan Kaufmann; 1992.
- (29) Itoh T, Minematsu N, Nakagawa S. Analysis of filled pauses and their use in a dialogue system. *The Journal of the Acoustical Society of Japan* 1999; **55**(5):333–342 (in Japanese).
- (30) Itoh T, Kitaoka N, Nishimura R. Subjective experiments on influence of response timing in speech dialogues. *IPSJ SIG Notes* 2008; **2008**(68):99–104 (in Japanese).

Ryota Nishimura (Non-member) was born in Mie, Japan, on May 20, 1982. He received his B.E. and M.E. degrees from Toyohashi University of Technology in 2005 and 2007, respectively, and is currently a Ph.D. student at Toyohashi University of Technology. His research interests include spoken dialog systems. He is a member of the Information Processing Society of Japan (IPSJ), the Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Japanese Society for Artificial Intelligence (JSAI).



Seichi Nakagawa (Member) was born in Kyoto, Japan, on November 18, 1948. He received his Ph.D. in Engineering from Kyoto University in 1977. He joined the faculty of Kyoto University in 1976 as a research associate in the Department of Information Sciences. He moved to Toyohashi University of Technology in 1980. From 1980 to 1983, he was an assistant professor, and from 1983 to 1990 he was an associate professor. Since 1990, he has been a professor in the Department of Information and Computer Sciences, at Toyohashi University of Technology. From 1985 to 1986, he was a visiting scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. He received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronics and Telecommunication Engineers. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He is a fellow of IPSJ.

