

# Prosody Change and Response Timing Analysis in Spontaneously Spoken Dialogs and Their Modeling in a Spoken Dialog System

Ryota NISHIMURA<sup>1</sup>, Norihide KITAOKA<sup>2</sup>, and Seiichi NAKAGAWA<sup>1</sup>

<sup>1</sup>Department of Information and Computer Sciences, Toyohashi University of Technology

<sup>2</sup>Graduate School of Information Science, Nagoya University

{nishimura, nakagawa}@slp.ics.tut.ac.jp, kitaoka@sp.m.is.nagoya-u.ac.jp

## Abstract

If a dialog system were to respond to a user as naturally as a human, interaction would be smoother. Imitating the human prosodic behavior of utterances is important in computer-human natural conversations. In this paper, to develop a cooperative/friendly spoken dialog system, we analyzed the correlations between F0 synchrony tendency or overlap frequency and subjective measures: “liveliness,” “familiarity,” and “informality” in human-human dialogs. We also modeled the properties of these features and implemented the model on our dialog system that generated the response timing of *aizuchi* (back-channel), turn-taking based on a decision tree in real time, and dynamical F0 changes to realize chat-like conversations.

**Index Terms:** spoken dialog system, prosody control, response timing

## 1. Introduction

Recently, computer performance has been augmented and automatic speech recognition (ASR) technology has also been improved. So ASR technology can be used in many situations. An ASR-based interface is greatly expected. Many spoken dialog systems have also been developed such as tourist information, information retrieval, car navigation, and so on. Since traditional systems, however, did not react to users during a user utterance, they could not know whether the system heard the utterance. Moreover, such systems responded with a somewhat ‘flat’ voice, producing a ‘stiff’ impression. In the future, spoken dialog systems are expected to sound more familiar to achieve smoother dialogs. Since human-human chat-like conversation is considered one ‘ideal’ smooth dialog, we are trying to make a spoken dialog system that can imitate various phenomena appearing in human-human dialogs to make the dialog system as natural as humans.

In Japanese human-human dialogs, such well-timed responses as ‘*aizuchi*’ (sometimes called ‘back-channel’) and turn-taking play important roles in smooth dialogs.

When making an ‘*aizuchi*’ or taking turns, humans usually pause for an appropriate length before talking, but sometimes they overlap their partner’s utterances. Such timing that includes overlaps is crucial in smooth dialogs.

In smooth and cooperative human-human conversations, such prosody as pitch is synchronized between speakers. According to Kakita [1], if a speaker’s F0 is high in a conversation, the other speaker’s F0 will also be raised in simple question-answer dialogs. Nagaoka et al. [2] showed that switching pause durations between dialog partners indicate a positive correlation. These suggest that appropriately controlling the prosody of the system’s response is necessary in spoken dialog systems to enable talk that is as smooth and natural as human-human dialogs.

Table 1: Dialog data in CSJ

ID	content	# of dialogs	time [hours]
D01	interview for SPS	16	3.4
D02	task dialog	16	3.1
D03	free dialog	16	3.6
D04	interview for APS	10	2.1

In this study, we first analyze human-human dialogs to learn how prosody interacts between speakers. Then we model the tracking tendency of F0 and implement the model as well as the response timing control to our dialog system.

## 2. Relation of prosody between speakers in human-human conversation

### 2.1. Dialog corpus

We investigated the prosodic features in human-human conversation by using the Corpus of Spontaneous Japanese (CSJ) provided by The National Institute for Japanese Language [3]. The corpus contains spontaneous speech in modern Japanese with additional information for research as well as 7.5 million words and 660 hours of voice. It is regarded as a resource for the research of automatic speech recognition (ASR). Most of the data are monologues such as Academic Presentation Speech (APS) and Simulated Public Speaking (SPS), but CSJ has also dialogs, as shown in Table 1.

We used these data for dialog analyse. CSJ has 58 conversations of about 10-20 minute durations each for a total of 12 hours.

In ‘interview for APS’ and ‘interview for SPS,’ interviewers asked many questions about their presentation. The two female interviewers are in their twenties and thirties.

In the task-oriented dialog, the earnings of Japanese entertainers is discussed. They first looked at a list of entertainers and then discussed the rank of their earnings. Finally they sorted them by rank. In the free dialog, no topic is provided *a priori*. Each dialog is about 10 minutes long. Sixteen pairs in each task consist of identical pairs.

These dialogs contained woman-woman dialogs and woman-man dialogs, but not man-man.

### 2.2. Change of fundamental frequency in a dialog

The fundamental frequency (F0) contours of a dialog speech in the corpus are shown in Figure 1. The difference of dotted types indicates different speakers (speakers L and R). Values are indicated by logarithmic scale (log F0) and normalized by subtracting the entire mean value of each speaker.

There are two sub-topics of conversation in the figure. In

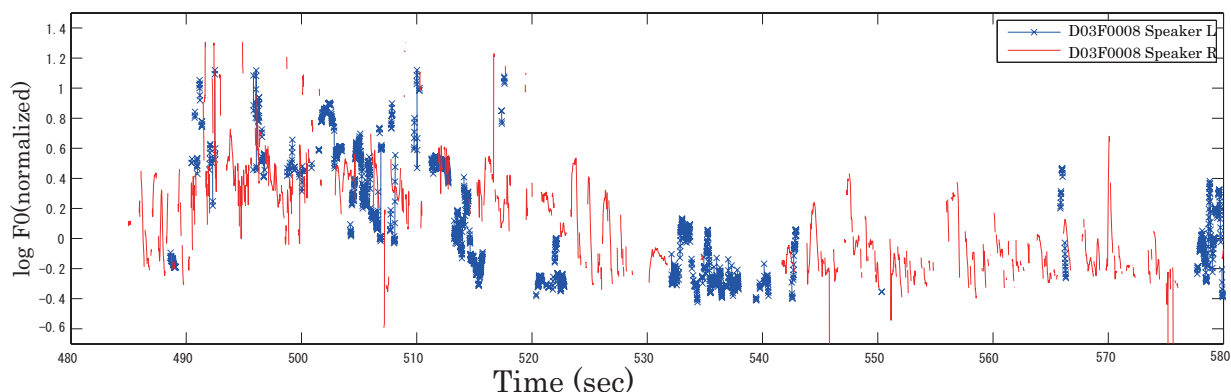


Figure 1: Example of F0 contours in a dialog in CSJ

Table 2: Correlations of log F0 between speakers in a dialog

ID	Max.	Min.	Avg.
D01	0.382	0.070	0.195
D02	0.477	-0.002	0.222
D03	0.521	0.012	0.234
D04	0.206	-0.005	0.085
avg.(all)	0.397	0.019	0.184
avg.(D02+D03)	0.499	0.005	0.228

the first half, they discussed ‘future dreams of children.’ One introduced a funny story and they often laughed and took many turns. The topic of the latter half was ‘children’s language acquisition,’ a relatively serious theme. Both speakers relaxed.

In the figure, the F0 value is higher in the “lively” region, and the dynamic range is also larger. On the other hand, in the “relaxed” region, the F0 value is not so varied from the mean value (near zero), and the dynamic range is also small.

These results indicate that “liveliness” relates to such prosodic features as F0. The prosody of the two speakers was probably influenced by each other.

### 2.3. Correlation of fundamental frequency between speakers in conversation

We investigate the correlation of fundamental frequency among speakers in dialogs.

We calculate the correlation of the utterance-wise log F0. We first determine the mean of all utterances and arrange them at the center of the utterance in the time axis. To associate the value with that of the conversation partner at the same time point, the partner’s values are linearly interpolated for the previous and following utterances.

The correlation values of the log F0 between two speakers are shown in Table 2. Positive correlation was observed in 54 dialogs out of 58 conversations. This indicates that the pitch of speech has synchrony tendencies based on the opposite speaker. The types of dialogs also affect the correlation value. In the free-style dialogs (D02 and D03), correlation values are higher than the interview-style dialogs (D01 and D04). This means that the log F0 of the two speakers synchronizes in the free-style dialogs better than in the interview-style. A gender difference is also observed. The correlation in the woman-woman dialogs was higher than in the other dialogs.

### 3. Relation between impression of dialog and prosodic phenomena

We investigate the relation between the impression when subjects listen to the dialogs and phenomena (such as F0 correlation between the two speakers, overlap frequency, and filler frequency) in the CSJ corpus. The four subjects answered questionnaires on the following items about each conversation after listening to the dialogs.

- Familiarity
  - familiar (5 4 3 2 1) hesitant
- Liveliness
  - lively (5 4 3 2 1) not lively
- Whether they agree with other speaker
  - agreeing (5 4 3 2 1) disagreeing
- Relative position in society of speaker L (age difference)
  - higher (5 4 3 2 1) lower
- Frankness of speaker L (interviewer)
  - careful (5 4 3 2 1) frank
- Frankness of speaker R (interviewee)
  - careful (5 4 3 2 1) frank
- Expression of speaker L (interviewer)
  - with many honorifics (5 4 3 2 1) without
- Expression of speaker R (interviewee)
  - with many honorifics (5 4 3 2 1) without

“Familiarity” and “liveliness” were evaluated considering their impressions of both speakers. “Relative position in society” evaluated whether speaker L seemed higher or lower than speaker R. “Frankness” denotes the impression received from each speaker’s utterances. “Expression” shows the usage of honorific expressions. Subjects evaluated whether the speaker often used honorifics in the dialog.

Four subjects (one male) randomly listened to various parts of all the sample data for about ten minutes to understand the mood and then listened to each sample for at least five minutes.

Here, we checked correlation between subjects to see individual differences of questionnaire results among subjects. The questionnaire result of “liveliness” is shown in Table 3. The average value is 0.470. The average results of other questionnaire items are shown in Table 4.

Then, we investigated the correlation between the questionnaire results and each dialog phenomena: “Overlap frequency,” Eq. (1), is the rate of overlapping responses in all turn-taking. “Filler frequency,” Eq. (2), is the rate of utterances with fillers in all utterances.

Table 3: Correlations of evaluation of liveliness between subjects

subject	correlation
subject 1 - subject 2	0.523
subject 1 - subject 3	0.538
subject 1 - subject 4	0.630
subject 2 - subject 3	0.314
subject 2 - subject 4	0.342
subject 3 - subject 5	0.476
average	0.470

Table 4: Average of correlations between subjects for each questionnaire item

questionnaire item	average of correlation
familiarity	0.444
liveliness	0.470
agree opinion	0.387
relative position of L	0.478
frankness of L	0.399
frankness of R	0.384
expression of L	0.300
expression of R	0.262

$$OverlapFrequency = \frac{\#OverlappingResponses}{\#TurnTaking} \quad (1)$$

$$FillerFrequency = \frac{\#Utterances\_with\_Fillers}{\#Utterances} \quad (2)$$

In Table 5, the overlap frequency generally indicates high correlation values. When overlap frequency is high, “familiarity,” “liveliness,” and “agreement” scores are also high. In fact, dialogs with many overlaps create a familiar and lively impression, and speakers tend to agree with their conversational partners. The negative values of “frankness” indicate that subjects get a frank impression from the dialog.

The correlation of F0 values is positively correlated with “familiarity,” “liveliness,” and “agreement.” These results mean that the F0s of the speakers synchronize well and that they often use overlaps. The dialogs seem familiar, lively, and frank, and they proceed with agreement between speakers.

Filler frequency positively correlates with “relative position in society” and “expression,” meaning that fillers and honorific expressions frequently happen when the conversational partner is older.

The relation of the overlap frequency and the subjective liveliness score is shown in Figure 2. A correlation is obvious.

Table 5: Correlations between questionnaire results and each dialog phenomenon

	correlation of F0	overlap frequency	filler frequency
familiarity	0.348	0.627	0.072
liveliness	0.350	0.718	0.127
agreeing with their opinions	0.279	0.638	0.090
relative position of L	-0.282	-0.098	0.267
frankness of L	-0.283	-0.417	0.108
frankness of R	-0.266	-0.637	0.047
expression of L	-0.340	-0.238	0.265
expression of R	-0.182	-0.404	0.289

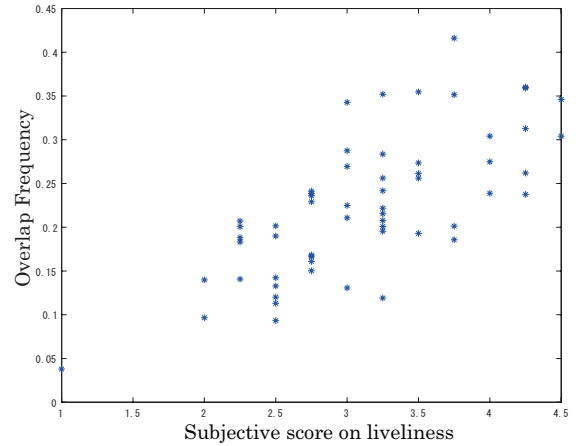


Figure 2: Plot of overlap frequency against subjective score of liveliness

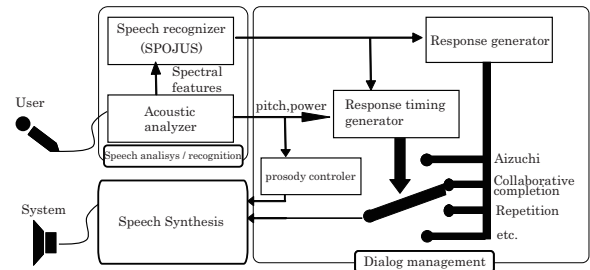


Figure 3: Schematic diagram of dialog system

## 4. Dialog system

### 4.1. Structure of dialog system

In Section 3, it is revealed that the speaker’s F0 and overlapping responses have strong relations to familiarity and liveliness. In this section, we explain the architecture of our spoken dialog system and how to implement prosody changes and overlapping responses into it.

The system is composed of three major parts, as shown in Figure 3. The acoustic analyzer and speech recognizer output the recognition hypotheses and the pitch/power contour patterns of the user utterances. The response generator, the timing generator, and the prosody controller generate response sentences, response timing, and the F0s of the response utterance, respectively, using the recognition results and the prosodic information of input utterances. Response sentences and F0s are sent to the speech synthesizer at the timing generated by the response timing generator.

### 4.2. Modeling of response timing

Previously we proposed a decision tree-based timing generator [4], but it can only treat the response at the end of user utterances. We modified this method to enable it to generate overlapping responses by scanning every segment whenever the user speaks. We used a decision tree to model response timing. But phenomena that rarely occur in the corpus cannot be learned. Therefore the rules for these phenomena are written by humans while referring to [5–7].

The RWC corpus [8] is used to train the decision tree, and C4.5 is used for machine learning. The following are the fea-

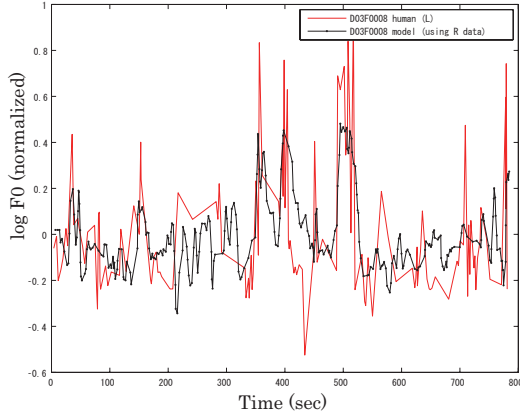


Figure 4: F0 value from dialog and output from model

tures of the decision tree:

- Duration from start time of user’s preceding utterance
- Elapsed time from end of previous user utterance
- Elapsed time from end of previous system utterance
- Pitch/power contour of the last 100 ms (consisting of three types)
- Pitch/power contour of the last 500 ms (consisting of five types)

The total number of features is 19. The decision tree outputs one “*aizuchi* (back-channel),” “turn-taking (system response),” “wait,” “repetition,” and “collaborative completion” every 100 ms.

#### 4.3. Modeling of F0 change

In human-human dialogs, the F0 change has a synchrony tendency between two speakers in lively dialog, as shown in Section 2. To achieve this tendency with the system, we model prosody control. The system monitors the user’s F0 and synchronizes it by following the user’s F0.

The base log F0 value of the utterance at turn  $t$ ,  $M(t)$ , is determined by the following equations:

$$M(t) = \mu_{sys} + \alpha_{sys}(t), \quad (3)$$

$$\alpha_{sys}(t) = \alpha_{sys}(t-1) + K(\alpha_{usrN\mu} - \alpha_{sys}(t-1)),$$

where  $\mu_{sys}$  is the standard (that is, average) F0 value of the system that does not change depending on time,  $\alpha_{sys}(t)$  offset of F0 value at turn  $t$ , and  $\alpha_{usrN\mu}$ , the average of the log F0 of the user’s last  $N$  utterances, which is the target value for the system.  $K$  is a time constant.

Here,  $K$  and  $N$  are set to 0.7 and 3 in the implementation, respectively.

To evaluate performance, we compared the output of the model by inputting one side of the corpus data with the log F0 sequence of the other side. The values are normalized by subtracting the mean value. Examples of the output of the model when inputting the speech of speaker R and the log F0 sequence of speaker L are plotted in Figure 4.

The average correlation between the CSJ corpus and the model output is shown in Table 6. Here, we only used D02 and D03 in Table 2. Positive correlations are observed, which are comparable to the correlation between the two speakers in Table 2. Notice that the initiative of the dialog must be considered. This model only chases the user’s log F0, and this strategy is only appropriate when the user has the initiative. So the model should be evaluated using the time periods in which the user

Table 6: Correlation between contours of real human and model in corpora D02 and D03. Top N means correlations obtained from only the dialog data having top N correlation between human speakers in the dialog

Data	Max.	Min.	Average
Top 11	0.553	0.361	0.451
Top 25	0.553	0.227	0.359
All	0.553	-0.025	0.206

has the initiative. Now we use all the periods in the corpus, and analysis of the corpus separated by the initiative is future work. Modeling the system’s behavior with initiative is also future work.

## 5. Conclusion

In this paper, we analyzed the relation between the synchronicity of prosody and liveliness in human-human dialogs and modeled it to develop a cooperative spoken dialog system. The dialog system based on this method has been developed [9]. “Liveliness” relates to such prosodic features as F0. The prosody of two speakers is probably influenced by each other. Some conversations contain a strong correlation of F0 between speakers. We investigated the relation between impressions when subjects listened to dialogs and prosodic phenomena. When the F0s of speakers synchronize well and they often use overlaps, the dialog seems familiar, lively, and frank, and proceeds with agreement between the speakers. To realize this synchrony in the dialog system, we proposed a model to chase user’s F0 and showed that it could simulate the F0 behavior of humans well.

Our prosody change model only passively follows users, but the system should actively change prosody depending on the dialog situation. In the future, we will subjectively evaluate this system. We will also compare subjective evaluation results using recorded human and synthesized voices as system output.

## 6. References

- [1] K. Kakita, “Inter-speaker interaction of F0 in dialogs,” *Proceedings of ICSLP-1996*, pp. 689–692, 1996.
- [2] C. Nagaoka, M. Komori, and S. Yoshikawa, “Synchrony tendency: interactional synchrony and congruence of nonverbal behavior in social interaction,” *Proceedings of Active Media Technology 2005 (AMT-2005)*, pp. 529–534, 2005.
- [3] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese,” *Proceedings of the Second International Conference of Language Resource and Evaluation*, vol. 2, pp. 947–952, 2000.
- [4] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, “Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems,” *Information and Media Technologies*, vol. 1, no. 1, pp. 296–304, 2006.
- [5] H. Noguchi and Y. Den, “Prosody-based detection of the context of backchannel responses,” *Proceedings of ICSLP-98*, pp. 487–490, 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] 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.
- [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. Nishimura, N. Kitaoka, and S. Nakagawa, “A spoken dialog system for chat-like conversations considering response timing,” *Proc. Text, Speech & Dialog TSD*, 2007. (to appear)