interspeech 2008 incorporating SST 08

Analysis of Relationship Between Impression of Human-to-human Conversations and Prosodic Change and Its Modeling

Ryota Nishimura[1], *Norihide Kitaoka*[2], *and Seiichi Nakagawa*[1]

[1]Department of Information and Computer Sciences, Toyohashi University of Technology
[2]Graduate School of Information Science, Nagoya University
{nishimura, nakagawa}@slp.ics.tut.ac.jp, kitaoka@nagoya-u.jp

Abstract

If a dialog system could respond to a user as naturally as a human, the interaction would be smoother. Imitating human prosodic characteristics of utterances is important in computer-to-human natural interaction. To develop a cooperative/friendly spoken dialog system, we analyzed the correlation between the fundamental frequency's synchrony tendency, or overlap frequency, and subjective measures of "liveliness", "familiarity", and "informality" in human-to-human dialogs. We also modeled the properties of these features to realize chat-like conversations in our spoken dialog system.

Index Terms: spoken dialog system, prosody control, response timing, chat-like conversation

10.21437/Interspeech.2008-160

1. Introduction

Recently, automatic speech recognition (ASR) technology has been improved and used in many situations due to advancements in computer performance. Many spoken dialog systems have been developed for such applications as tourist information, information retrieval, and car navigation. However, since traditional systems do not noticeably react to a user's voice, a user cannot tell whether the system heard his or her utterance or not. Moreover, such systems respond with a somewhat 'flat' voice and produce a '*stiff*' impression. In the future, spoken dialog systems are expected to sound more natural to achieve smoother dialogs. Since human-to-human chat-like conversation is considered as the 'ideal', we have been trying to make a spoken dialog system that can imitate various phenomena characteristic in human-to-human dialogs [1].

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

During '*aizuchi*' or turn-taking, humans usually pause for an appropriate length before talking, but they sometimes overlap their partner's utterances, i.e., barge-in. Such timing that includes overlaps is crucial in smooth dialogs.

In smooth and cooperative human-to-human conversations, the prosody as pitch is synchronized between speakers. According to Kakita [2], if a speaker's fundamental frequency (F0) is high in a conversation, the other-side speaker's F0 will also be raised in simple question and answer dialogs. Nagaoka et al. [3] showed that switching pause durations between dialog partners indicate a positive correlation. These suggest that to enable communication as smooth and natural as human-to-human dialogs it is necessary to appropriately control the prosody of a system's response.

We first analyzed human-to-human dialogs to find how prosody interacts between speakers. Then we modeled the tracking tendency of F0, power, and speech rate.

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

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

2.1. Dialog corpus

We investigated the prosodic features in human-to-human conversation by using the Corpus of Spontaneous Japanese (CSJ) provided by The National Institute for Japanese Language [4], which we previously investigated [5]. 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 research in ASR. Most of the data are monologue form such as the Academic Presentation Speech (APS) and Simulated Public Speaking (SPS), but CSJ also has other dialogs as shown in Table 1.

We used these data for dialog analyses. 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 to a person about their presentation. The two female interviewers were in their twenties and thirties.

These dialogs contained female-to-female and female-to-male dialogs, but not male-to-male.

We divided each dialog into 709 topics, each about one minute, to perform a detailed analyses.

2.2. Change of fundamental frequency in a dialog

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

There were two sub-topics of conversation in the figure. In the left half, they discussed 'future dreams of children.' One introduced a funny story and they often laughed and took many turns. The topic of the right half was 'children's language acquisition', a relatively serious theme, and both speakers not lively.

The F0 value was higher in the "lively" region, and the dynamic range is also larger. On the other hand, in the "not lively" region, the F0 value was not so varied from the mean value (near zero), and the dynamic range was also small.

These results indicate that "liveliness" is related to such prosodic features as F0. The prosody of the two speakers was

Accepted after peer review of full paper

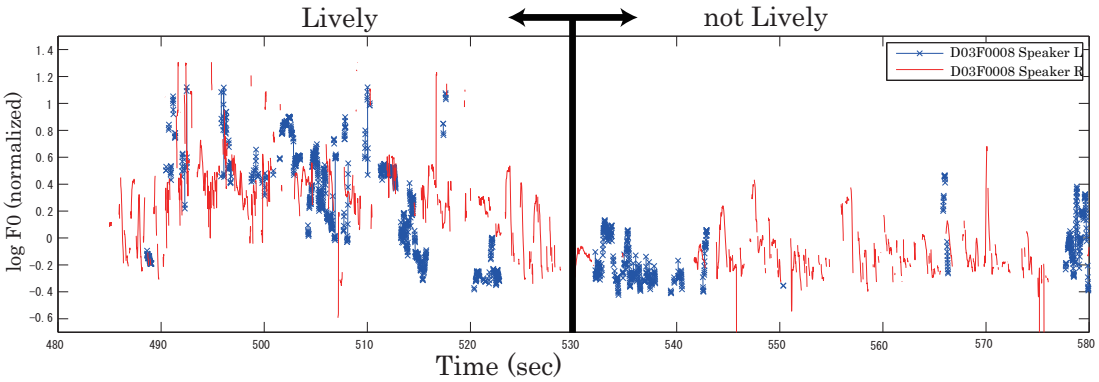

Figure 1: Example of F0 contours in a CSJ dialog

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

ID	Max.	Average	StdDev	Positive(%)
D01	0.716	0.145	0.247	70.6
D02	0.758	0.202	0.293	73.2
D03	0.710	0.166	0.265	72.0
D04	0.771	0.047	0.288	52.2
avg.(all)	0.771	0.150	0.276	68.7

Table 3: Percentage of dialogs that have a significant correlation in F0 ($p<0.05$)

Liveliness level	Dialog		Topic	
	%	Average correlation	%	Average correlation
> 4	100	0.247	100	0.345
> 3.5	100	0.246	45.8	0.229
> 3	91.4	0.228	45.6	0.143

probably influenced by each other.

2.3. Correlation of fundamental frequency between speakers in a conversation

We investigated the correlation of F0 between speakers in dialogs. The investigated dialog data was divided into 709 topics, each about one minute. The total number of topics was 709. Here, the topic in which the number of utterances from a speaker was less than 10 was excluded from the investigating data because it could not be analyzed correctly. Thus, the actual number of data was 566.

We calculated the correlation between two speaker's utterance-wise log F0. We first determine the mean of all utterances and arranged them at the center of one utterance in the time axis. To assign the value with that of the conversation partner at the same time point, the partner's values were linearly interpolated by using the previous and following utterances.

The correlation values of the log F0 between two speakers are shown in Table 2. A positive correlation was observed in 389 out of 566 topics (68.7%). This indicates that the pitch of speech had synchrony tendencies based on the other speaker. The types of dialogs also affected the correlation value. In the free-style dialogs (D02 and D03), the correlation values were higher than the interview-style dialogs (D01 and D04). This means that the log F0 of the two speakers was more synchronized in the free-style dialogs than in the interview-style. A gender difference was also observed. The correlation in the female-to-female dialogs was stronger than in the other dialogs (67% of the top 100 topics with a strong F0 correlation being female-to-female).

Table 3 shows the percentage of dialogs that have a significant correlation in F0 ($p<0.05$) [1] with the "liveliness" levels evaluated in Section 3. We can find that there is relation between the liveliness level and the correlation of F0 among speakers. Very lively conversations had significant correlations. There were significant correlations in all the dialogs with liveliness level greater than 3 in 5-point scale. Similary there were significant correlations in all the topics with liveliness level greater than 4.

3. Relation between impression of dialog and prosodic phenomena

We investigated the relation between the impression from the subjects when they listened to the dialogs and prosodic phenomena (such as F0, power, speech rate, and correlation between the two speakers) in the CSJ corpus. The six subjects (four males, two females) answered questionnaires on the following items on 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 the other speaker or not
  - agreeing (5 4 3 2 1) disagreeing
- Age difference
  - smaller (5 4 3 2 1) larger
- Frankness of speaker L (interviewer)
  - frank (5 4 3 2 1) careful
- Frankness of speaker R (interviewee)
  - frank (5 4 3 2 1) careful
- Expression of speaker L (interviewer)
  - without (5 4 3 2 1) with many honorifics
- Expression of speaker R (interviewee)
  - without (5 4 3 2 1) with many honorifics

"Familiarity" and "liveliness" were evaluated considering the impressions from both speakers, "Age difference" was evaluated between speakers L and R, "frankness" denotes the impression received from each speaker's utterances, and "expression" shows the usage of honorific expressions. Subjects evaluated each other on whether they often used honorifics in the dialog or not.

For divided topic data, we used two topics from each dialog. The total number was 116. The questionnaire items for topic-units were the same as for the dialog-units, and the item "synchronization" was added.

We checked the correlation between subjects to see individual differences of questionnaire results. The average results

[1] The null hypothesis of no correlation is rejected at the significant level of 5%

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

Questionnaire item	Average of correlation	
	Dialog	Topic
Familiarity	0.444	0.361
Liveliness	0.470	0.446
Agree opinion	0.387	0.241
Age difference	0.478	0.376
Synchronization	no result	0.337
Frankness of L	0.399	0.128
Frankness of R	0.384	0.178
Expression of L	0.300	0.134
Expression of R	0.262	0.159

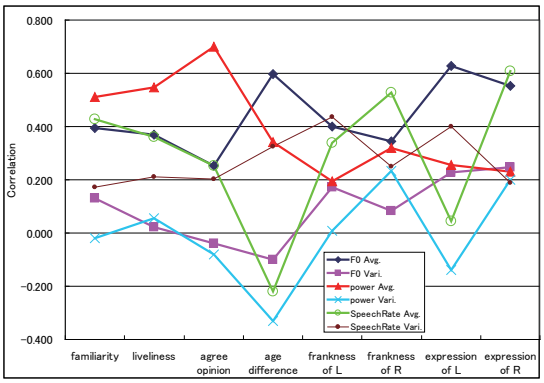

Figure 2: Correlation between questionnaire results and each dialog phenomenon

of other questionnaire items are shown in Table 4. The "dialog" was roughly a 10-minute conversation, and the "topic" was roughly a 1-minute conversation.

"Familiarity", "liveliness", "age difference", and "synchronization" could be stably evaluation and were strongly correlated. However, "frankness" and "expression" could not be stably evaluated.

Figure 2 shows the correlation between questionnaire values and each dialog phenomenon (F0 average/variance correlation, power average/variance correlation, speech rate average/variance correlation).

The correlation of F0 average values was strongly and positively correlated with "familiarity", "liveliness", "agreement," and "age difference". The correlation of power average values was positively correlated with "familiarity", "liveliness", and "agreement." The correlation of speech rate average value was positively correlated with "familiarity" and "liveliness". These results mean that when the prosodic changes of the speakers synchronized well with each other, then the dialogs seemed familiar, lively, frank, and in agreement.

4. Modeling of prosodic change

In human-to-human dialogs, the prosodic change (such as F0, power, speech rate) has a synchronic tendency between two speakers in a lively dialog as shown in Section 2. To achieve this tendency with the system, we modeled a prosody control. The system monitors the user's prosodic change and synchronizes its own change by following the user's prosodic change.

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

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

where μ_{sys} is the standard (average) F0 value of the system that

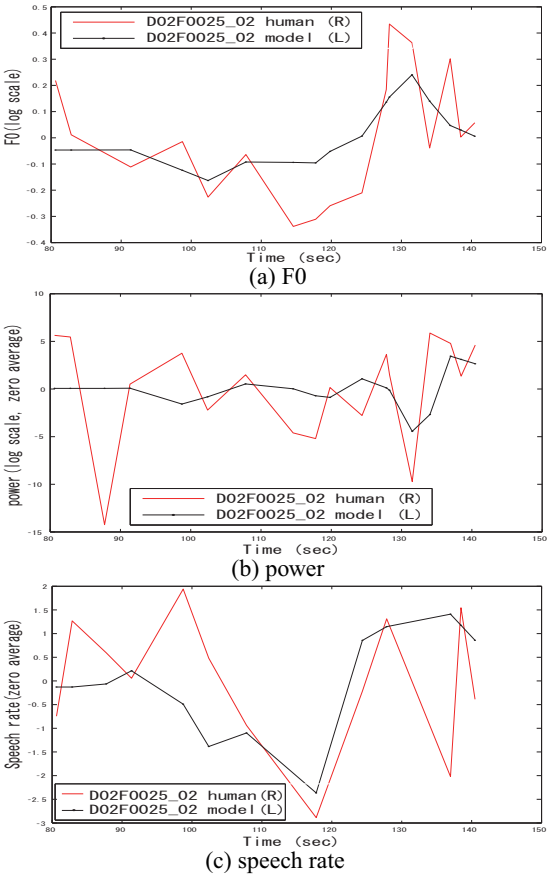

Figure 3: Prosodic value from dialog and output from model

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

To evaluate performance, we compared the output of the model by inputting one side speaker utterances of the corpus with the log F0 sequence of the other side speaker utterances. The values were normalized by subtracting the mean value. Examples of the output of the model when input of the speech of speaker R and the log F0 sequence of speaker L were plotted are shown in Figure 3. It shows that the model can be used for accurately imitating human prosodic change.

We experimented with varying the model. N was changed from 1 to 5 and K from 0.3 to 1.0. The best values for each model are indicated in Table 5.

The average correlations between the CSJ corpus and the model output are shown in Table 5. Positive correlations were observed, which were comparable to the correlation between the two speakers in Table 2. Notice that the initiation of the dialog must be considered. This model only imitates the user's log F0, and this strategy is only appropriate when the user initiates the conversation. So the model should be evaluated using the time periods in which the user initiates. We used all the periods in the corpus, and the analysis distinguished by the initiatives and the modeling based on the analysis are future works.

As a result, for the topic-unit, F0 should make the model slowly imitate (i.e., K=0.4) only the utterance immediately before (i.e., N=1). The power should make the model quickly imitate (i.e., K=1.0) the utterance immediately before (i.e., N=1). The speech rate should make the model slowly imitate (i.e., K=0.4) the past several utterances (i.e., N=5).

When the model parametor N is set to 1, it means that the model does not use any context histories. Table 5 shows that op-

Table 5: Correlation for best model parametor values (N, K)

(a) Dialog-unit

	N	K	Max.	Average	StdDev.
F0	1	1.0	0.663	0.185	0.141
Power	2	0.8	0.479	0.129	0.128
Speech rate	2	1.0	0.245	-0.137	0.148

(b) Topic-unit

	N	K	Max.	Average	StdDev.
F0	1	0.4	0.899	0.139	0.315
Power	1	1.0	0.823	0.103	0.281
Speech rate	5	0.4	0.987	-0.082	0.315

Table 6: Correlation for best model parametor values (N, K) for every topic-unit

	Max.	Average	StdDev.
F0	0.899	0.338	0.252
Power	0.823	0.294	0.242
Speech rate	0.987	0.173	0.254

timal Ns in the models for the power and the speech rate were not 1 but 2 in dialog-wise modeling. In topic-wise modeling, optimal N for the power model was 5. From these results, context histories are effective to model the prosodic change in some cases.

Meanwhile, When the model parametor K is set to 1.0, it means that the model has no time delay to catch up with the user's prosodic change. Table 5 shows that optimal K for the power model was 0.8 in dialog-wise modeling, and those for the F0 and the speech rate models was 0.4 in topic-wise modeling. These results show that the use of time constant K is also effective to model the prosodic change.

Table 5(b) shows strong positive correlation, which is comparable to the correlation between the two speakers in Table 2.

Table 6 shows the results of selecting the value of a model variable at every topic. The averages are higher compared with Table 5 (b). However, the standard deviation was lowered.

This suggests that when the model variable is selected according to the situation, this model can be used to control the prosodic change similar to humans.

5. Effect degree to conversational impression of prosodic information

In human-to-human dialogs, the prosodic change (such as F0, power, speech rate) had a synchronic tendency between two speakers in a lively dialog as shown in Section 2. And there is a correlation between the impression of the conversation and the prosodic change as shown in Section 3. In accordance with these phenomena, we modeled the prosodic change.

However, whether humans gather impressions of conversations only by the prosodic information is not known. Therefore, we created audio-data that preserves only the prosodic information of speech (hereafter called "humming-speech"), and subjects were evaluated on their impression of such speech.

Three subjects (one male, two females) answered questionnaires on the items (items are same in Section 3). The item "Expression" is a linguistic feature, so it was excluded.

The average results of the correlation between the subjects' humming-speech evaluation are shown in Table 7. Some items

Table 7: Average correlation of humming-speech evaluation result between subjects

	Dialog	Topic
Familiarity	0.347	0.505
Liveliness	0.425	0.584
Synchronize	no result	0.446

Table 8: Correlation between normal speech evaluation and humming-speech evaluation (two of three subjects average)

	Dialog	Topic
Familiarity	0.398	0.492
Liveliness	0.457	0.400
Synchronize	no result	0.357

(such as "agree opinion", "age difference", and "frankness") were difficult to evaluate based on impression, so they were excluded from the table.

Table 7 shows that the short conversation data (topic-unit) had a stronger correlation than that of the long conversation data (dialog-unit). It is thought the reason is that the topic-unit was shorter than the dialog-unit, so the difference between the subjects' evaluation decreased. The correlation of the topic-unit was lower than that of the dialog-unit in Table 4 because the language information from a dialog-unit can be used for evaluation in Table 4, but language information cannot be used from a topic-unit because it is too short, and the correlation weakens.

The correlation between normal speech and humming-speech evaluations is shown in Table 8. Here, the correlation of one subject was much weaker than that of the other two subjects because he could not perform the task, so he was excluded from the statistics. Table 8 shows that the correlation value increased from 0.3 to 0.5.

It was shown that the impression signifiers of the conversation (such as "familiarity", "liveliness" and "synchronize") were able to be evaluated with only the use of prosodic information.

6. Conclusion

We analyzed the relationship between the synchronicity of prosody and liveliness in human-to-human dialogs and modeled it to develop a cooperative spoken dialog system. "Liveliness" is related to such prosodic features as F0, power, and speech rate. The two speakers' prosody is probably influenced by each other. Some conversations contain a strong correlation of prosodic change between speakers. We investigated the relationship between subjects' impressions when they listened to dialogs and prosodic phenomena. When the prosodic change of speakers synchronized well, it seemed the dialog became familiar, lively, and frank, and the speakers were in agreement. To realize this synchrony in the dialog system, we proposed a model to imitate a user's F0/power/speech rate and showed that it could accurately simulate the prosodic characteristics of humans.

Our prosodic 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.

7. References

[1] R. Nishimura, N. Kitaoka, and S. Nakagawa, "A spoken dialog system for chat-like conversations considering response timing," *TSD 2007*, pp. 599–606, 2007.

[2] K. Kakita, "Inter-speaker interaction of F0 in dialogs," *Proceedings of ICSLP-1996*, pp. 689–692, 1996.

[3] 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.

[4] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," *Proceedings of SSPR 2003*, pp. 7–12, 2003.

[5] R. Nishimura, N. Kitaoka, and S. Nakagawa, "Prosody change and response timing analysis in spontaneously spoken dialogs and their modeling in a spoken dialog system," *Proceeding of the Interspeech 2007*, pp. 2565–2568, 2007.