PREDICTION OF F0 PARAMETER OF CONTEXTUALIZED UTTERANCES IN DIALOGUE

Y. Yamashita *R.Mizoguchi*

I.S.I.R., Osaka University
8-1, Mihogaoka, Ibaraki-shi, Osaka, 567 JAPAN
{yama,miz}@ei.sanken.osaka-u.ac.jp

ABSTRACT

In order to synthesize natural spoken dialogue, it is necessary to incorporate dialogue information into generation of the surface sentence and the prosody. This paper describes the prediction of F0 maximum for minor phrases in dialogue based on a two-step predictive method. Special attentions are directed to specific phrases containing the person's name or the day of the week in the schedule arrangement task in order to narrow the diversity of characteristics of F0 parameters in dialogue. Seven features were identified as dialogue information which are useful to predict the F0 parameter. Two D-rule sets derived from the person's name or the day of the week are very similar to one another. They reduce the total prediction errors by about 50% for the data which have much influence of dialogue context.

1. INTRODUCTION

Calculation of prosodic parameters requires various information in spoken language synthesis for dialogue systems. Prosody rules in the text-to-speech (TTS) system are not sufficient for generation of spoken dialogues. Prosody prediction based on the dialogue information must be incorporated into the spoken dialogue synthesis system. Prosodic characteristics of utterances in a dialogue vary in accordance with the dialogue context even if their surface sentences are the same. Relationship between prosody and dialogue context is a crucial issue of spoken dialogue synthesis. Many studies have been conducted for qualitative intonational prediction from dialogue information, especially for placement of the pitch accent for English sentences[1][2]. The quantitative prediction is very important as well as the qualitative one, especially for F0 contour of Japanese sentences. The authors have been proposed a method of quantitative prediction of F0 parameters from dialogue information and described evaluation of rules derived from utterance data for a task of the route inquiry[3]. However, diversity of characteristics of prosody in actual dialogues spreads over very wide range and the derived rules could not satisfactorily capture the behavior of the prosodic parameters. In this paper, we use more specific utterance data in a small task of schedule arrangement to predict F0 maximum of minor phrases in the dialogue utterance based on the two step predictive method.

2. TWO STEP PREDICTION OF PROSODY IN DIALOGUE

2.1. Framework

The authors have proposed a method of modeling contextual effects on prosodic parameters in dialogue based on the two-step prediction. In this framework, the prosodic parameter is predicted by two sets of rules, as depicted in Figure 1. The first set of rules is composed of conventional rules for TTS and this set is called S-rule. (Note that this fundamental rule set was denoted by the T-rule in our former paper[3].) The S-rule predicts the prosodic parameters using syntactic or lexical features of the surface sentences. Then, the second set of rules, D-rule, adjusts the prosodic parameters to dialogue context using higher level information, such as topic, utterance type, dialogue history, and so on.

The generation process of S-rule and D-rule is depicted in Figure 2. The S-rule is derived from isolated utterances based on a stochastic modeling technique, and it is applied to utterances in dialogues. The S-rule captures prosodic characteristics of isolated utterances very well. However, the S-rule knows nothing about how to predict prosody appropriate to the dialogue context. Most errors in predicting prosody for the dialogue utterances are caused by the contextual effects of the dialogue. The D-rule is derived from descriptions of the prediction error in terms of dialogue features describing higher level information which is not considered in the S-rule modeling. The D-rule is modeled by a stochastic technique again.

This two-step modeling has the following advan-

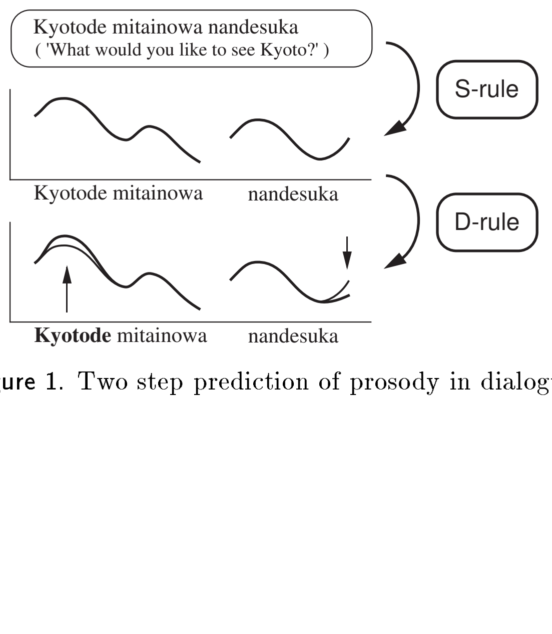

Figure 1. Two step prediction of prosody in dialogue

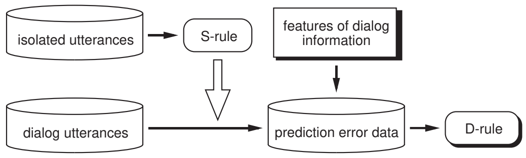

Figure 2. Flow of the S- and D-rule generation

tages.
(1) It can efficiently reuse conventional prosody rules of text-to-speech conversion.
(2) It makes contextual effects clear which are dependent on the dialogue context.
(3) The number of free parameters in a stochastic modeling is reduced for predicting prosodic parameters in a dialogue. The prosodic parameters are partially predicted by the S-rule and only the dialogue context effects remain. This is very important in cases where small number of dialogue data are available for quantitative prediction.

2.2. S-rule Modeling

In this paper, the two sets of rules predict F0 maximum of each minor phrase of utterances in the two-stage F0 control model, which Abe and Satoh have proposed for isolated Japanese sentences[4]. In their paper, the 'global model' predicts this F0 parameter using the linear regressive method. We used the same stochastic modeling technique as theirs to obtain the S-rule. Each minor phrase was described in terms of 7 syntactic features. They include dependency relationship from the preceding and to the following phrases, accent type of the preceding, the current and the following phrases, the syllable count and part of speech of the current phrase.

2.3. D-rule Modeling

The D-rule models the change of the prosodic parameter by the dialogue context using the linear regression method, too. A training case for D-rule modeling corresponds to an each minor phrase in dialogue utterances. We investigated characteristics of the dialogue context for the schedule arrangement task, and came up with a set of 7 features reflecting the context. The followings describe all the features of higher level information and their values.

[DF1] the semantic role of the utterance to the preceding utterance is classified into 8 categories: topic change, topic return, topic continuation, topic elaboration, positive/negative supplement, or response for request/WH-question.

[DF2] "yes" if the phase is followed by a coordinate phrase, "no" otherwise.

[DF3] "yes" if the phrase is positioned at the beginning of the utterance, "no" otherwise.

[DF4] the preceding category of the phrase is classified into 5 categories: normal word, filled pause, pause, disfluency, or beginning of the utterance.

[DF5] "yes" if the phrase is negligible, "no" otherwise.

[DF6] "yes" if a contrastive phase exists in the dialogue history, "no" otherwise.

[DF7] "yes" if the phrase is included in a utterance which lists similar information to the topic of the preceding utterances, "no" otherwise.

All minor phrases in an utterance have the same value for the *DF1* and the *DF7*, because the *DF1* and the *DF7* are features assigned not for each minor phrase but for an utterance. Other features, in general, have different values in an utterance. The training cases are described in terms of a prediction error value and these 7 features of higher level information.

3. EXPERIMENTS

3.1. Speech Data

Our framework of the two-step prediction of prosodic parameters in dialogue requires two kinds of speech data: isolated utterances which are independent of dialogue context and dialogue utterances which are contextualized in dialogue.

The material for isolated utterances is a set of 503 phonetically balanced sentences designed at ATR. These sentences were once uttered by a male non-professional speaker. F0 patterns were corrected by hand and the maximum F0 frequencies for the minor phrase were manually identified. The minor phrase is a unit of training cases because the maximum F0 is assigned to each minor phrase in Abe's F0 model. The number of cases in the isolated utterances was 3,288.

In order to collect contextualized utterances, dialogues between a questioner and an answerer were recorded with the task of schedule arrangement. The goal of a questioner in dialogue is to find unoccupied time of participants of two meetings which he will join. The questioner knows his own schedule and the participants of each meeting, but he knows nothing about the other participants' schedule. On the other hand, the answerer knows the schedule of all participants of the meetings except for the questioner and collaborates with the questioner. In this paper, we focus on the Japanese noun phrases containing proper nouns such as person's name or the day of the week in the schedule arrangement task, because these phrases have very important roles in the schedule arrangement task. Phrases containing the person's name or the day of the week in the utterances only of answerers were separately used for the D-rule modeling. The number of the answerer speaker is 7. Each answerer carried out 3 dialogues under different schedule settings. All answerers are male non-professional speakers and are not the same as the former speaker of the isolated utterances. The total number of dialogues is 21. The maximum F0 frequencies for minor phrases were also manually identified for the dialogue utterances. These dialogue utterances totally included 112 and 138 phrases of the person's name and the day of the week, respectively.

Table 1. Evaluation of S-rule

	training data	test data	averaged error [Hz]
(1)	isolated	isolated	10.0
(2-I)	isolated	dialogue-PN	18.8
(2-II)	isolated	dialogue-DW	18.8

Table 2. Evaluation of D-rule

(a) dialogue-PN

	test data	averaged error [Hz]		
		only S-rule	S- & D-rule	
		open	closed	open
(1)	all	18.8	11.7	13.7
(2)	E20*	33.9	15.9	18.6

(b) dialogue-DW

	test data	averaged error [Hz]		
		only S-rule	S- & D-rule	
		open	closed	open
(1)	all	18.8	12.5	14.1
(2)	E20*	34.4	15.8	17.7

(*E20: the cases of which the prediction error by the S-rule is larger than 20Hz.)

3.2. Evaluation of S-rule

Before discussing total prediction result of the F0 parameters, let us evaluate performance of the S-rule. Table 1 shows prediction errors by the S-rule. The S-rule was derived from the 503 isolated utterances. Table 1 (1) indicates the fundamental performance of the S-rule. The S-rule was evaluated for the isolated utterances based on 10-fold cross validation.

The prediction error by the S-rule increases to 18.1 Hz from 10.0Hz for both dialogue data. The dialogue-PN, -DW in Table 1 refers the phase data containing the person's name and the day of the week, respectively. This differences between (1) and (2) were mainly caused by the contextual effects of the dialogue.

3.3. Evaluation of D-rule

The prediction error data which the S-rule resulted for minor phrases in dialogues become a new set of training cases for learning the D-rule. All minor phrases were manually described in terms of new 7 features, *DF1*, *DF2*, ..., *DF7*. The D-rule was derived from these error data with the higher level information. The dialogue-PN and -DW were separately used for learning the D-rule.

The top rows (1) in Table 2 (a) and (b) show the total prediction error of F0 maxima over all minor phrases of the dialogue data. The prediction without the D-rule is open evaluation because the S-rule was derived from isolated utterances. The open evaluation with the D-rule was carried out by using 10-fold cross validation. This table shows that the D-rule reduced the total prediction errors by about 25% for both dialogue data.

Table 3. Partial correlation efficients of the D-rule

feature	partial correlation efficient	
	dialogue-PN	dialogue-DW
DF1	0.28	0.52
DF2	0.46	0.38
DF3	0.40	0.39
DF4	0.25	0.24
DF5	0.34	0.34
DF6	0.23	0.14
DF7	0.27	0.23

All utterances or all minor phrases do not necessarily have contextual effects of the dialogue. We can expect that cases which have large prediction errors by the first S-rule application are much influenced by the dialogue context. The D-rule was again evaluated only using the cases which have initial prediction errors larger than 20Hz. Sixty and Sixty three cases were matched with this condition for the dialogue-PN and -DW, respectively. Note that the D-rule learning was carried out using all the cases in each set of the dialogue utterances. The bottom rows (2) in Table 2 (a) and (b) show the second evaluation. The D-rule decreased the total prediction errors into about a half.

4. DISCUSSION

Table 3 shows the partial correlation coefficients of each feature for two D-rule sets. A feature of a large partial correlation coefficient makes a lot of contribution to predicting the target parameter. The *DF1*, *DF2*, *DF3*, and *DF5* have rather large weight, but the other features are not trivial. The multiple correlation coefficient of the linear regressive method indicates the strength of relationship between a target parameter and a set of features of the case. The D-rule sets score 0.73 and 0.66 for the dialogue-PN and -DW, respectively. These scores suggest the possibility of further improvement of the prediction accuracy using other features, though 7 higher level features capture the characteristics of the F0 parameter well.

Figure 3 depicts details of model parameters of the linear regressive method for the D-rule sets derived from two kinds of dialogue data. In this figure, we can find the same tendencies in two D-rule sets from dialogue-PN and -DW. Roughly speaking, each category for most of the features has very similar weights for two kinds of dialogue data. We tried to apply each D-rule set to the other dialogue data, which is complete open evaluation of the D-rules. Table 4 indicates that this evaluation on the different dialogue data does not increase prediction errors. Two D-rule sets derived from the person's name and the day of the week are very similar in this task.

As for the detailed observation of the category weights, however, there are some discrepancies especially for the *DF1*. In the task of the schedule arrange-

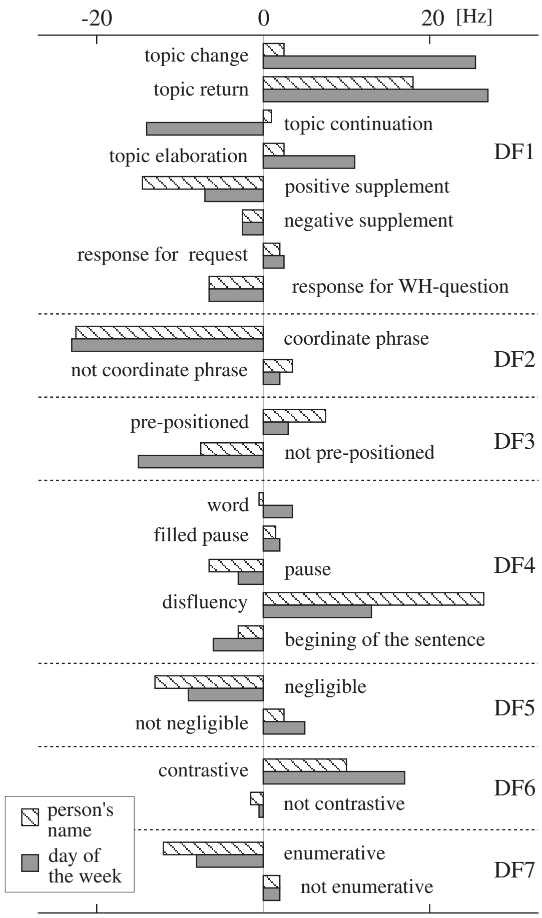

Figure 3. Model parameters of D-rule

Table 4. Crossed evaluation of the two D-rule sets

test data	training data	
	dialogue-PN	dialogue-DW
dialogue-PN	13.7	13.7
dialogue-DW	14.6	14.1

ment we adopted in this study, the participants of the meetings were given to both the questioner and the answerer, and their goal in a dialogue is to find a common unoccupied time. The change of the time or the day of the week was important for them, while the change of person's name did not give them so much information. This characteristic of the dialogues resulted in the large boost of 'topic change' for the dialogue-DW data. On the other hand, its weight for dialogue-PN is rather neutral.

The 'topic continuation' for *DF1* shows another large disagreement between two D-rule sets. The day of the week often appeared in phrases such as '*suiyoubi no juuji*(on Wednesday at 10 o'clock)' in dialogues. In the case that the utterance was classified into 'topic continuation', the focus was put on the time and F0 for the word representing the time was boosted. It gave the relative suppression of F0 for the day of the week.

The *DF2* has the large negative weight for the 'coordinate phrase'. This result is due to the incompleteness of the S-rule. The 'coordinate phrase' was mainly labeled to the word B in phrases such 'A *to* B (A and B)'. The suppression of F0 of the second or later accentual phrases is well known as the downstep phenomenon. A feature for coordination was not necessary in the S-rule modeling in this paper because ATR 503 phonetically balanced sentences, which was the training data for the S-rule, contain few coordinative expressions. Properly speaking, this feature should be considered in the S-rule modeling.

5. CONCLUSIONS

This paper described the prediction of F0 maximum for the minor phrase in dialogue based on the two-step predictive method. The diversity in dialogue tends to make it difficult to predict prosodic parameters when the wide range of context is scoped. In this paper, special attentions were directed to specific phrases containing the person's name or the day of the week in the schedule arrangement task. Seven features were identified as higher level information which was not considered in the fundamental S-rule modeling. They are very effective to predict the F0 parameter in dialogue and reduced the total prediction errors by about 50% for the data which have much influence of dialogue context.

The D-rule modeling separately generate rule sets for two kinds of dialogue data: the person's name and the day of the week. Two D-rule sets are very similar to one another. It means that the concepts of the person's name and the day of the week have very similar roles in the dialogue task of the schedule arrangement.

How to extract higher level information, such as the semantic role of the utterance, through a dialogue is a crucial issue but is not discussed in this paper. It will be future work along with the categorization of higher level information.

REFERENCES

[1] J. Hirschberg: "Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech", *Proc. of ESCA Workshop on Speech Synthesis*, Autrans, pp.181-184 (1990).

[2] A.I.C. Monaghan: "Intonation Accent Placement in a Concept-to-Dialogue System *Proc. of 2nd ESCA/IEEE Workshop on Speech Synthesis*, New York, pp.171-174 (1994).

[3] Y.Yamashita and R.Mizoguchi: "Modeling the Contextual Effects on Prosody in Dialog", *Proc. of EUROSPEECH '95*, 2, pp.1329-1332 (1995).

[4] M. Abe and H. Sato: "Two-Stage F0 Control Model Using Syllable Based F0 Units", *Proc. of ICASSP '92*, 2, pp.53-56 (1992).