MODELING THE CONTEXTUAL EFFECTS ON PROSODY IN DIALOG

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

This paper proposes a new method of stochastic modeling of prosodic contextual effects in dialog for the purpose of generating natural spoken language in dialog systems. F0 parameters are quantitatively predicted by two sets of rules; T-rule and D-rule. The T-rule is a conventional rule set for text-to-speech conversion. It models characteristics of isolated utterances without dialog context. The D-rule compensates the T-rule by modifying F0 parameters in accordance with the dialog context after application of the T-rule. The D-rule was inductively learned from the prediction error data which the T-rule resulted for the dialog utterances. The linear regressive method modeled the prediction errors with 7 dialog features manually annotated. The evaluation results of the D-rule are also shown for dialog utterances recorded under two different conditions.

1. INTRODUCTION

Prosodic characteristics of utterances in a dialog vary in accordance with the dialog context even if their surface sentences are the same. Relationship between prosody and dialog context is a crucial issue of spoken dialog generation. Many studies have been conducted for intonational prediction from dialog information, especially placement of the pitch accent for English sentences[1][2]. This qualitative prediction mainly addressed determination of features of dialog information, such as intention, given/new, focus, and so on, and categorization of their values.

Prosodic parameters must be calculated from text and dialog information for spoken language response in dialog systems. The quantitative prediction is very important as well as the qualitative one, especially for F0 contour of Japanese sentences. It has been also pointed out that generating F0 contour of an isolated sentence is modified when it is contextualized in a paragraph[3]. Relationship between changes of the prosodic parameter and the dialog information has been analyzed[4]. These findings were obtained by using monolog speech data. In order to realize natural spoken response in dialog systems, it is necessary to develop a framework of quantitative prediction of prosodic parameters and to incorporate dialog information in prosody generation. This paper proposes a new paradigm for quantitative modeling of the relationship between F0 parameters and the dialog information.

Noticeable differences between prosodic parameters of read and spontaneous speech in human-machine dialogs have been reported from a viewpoint of speech recognition[5]. However, we are interested in the difference between isolated and contextualized utterances rather than that of such speaking styles. This paper focusses on prediction of contextual effects on prosody from a viewpoint of spoken dialog synthesis.

2. A METHOD OF PREDICTING PROSODY OF SPOKEN DIALOGS

2.1. Total Framework

This section describes a method of modeling contextual effects on prosodic parameters in dialog. The prosodic parameter of utterances in a dialog is predicted by two sets of rules. The first set of rules is composed of conventional rules for text-to-speech (TTS) conversion and this set is called T-rule. The T-rule predicts the prosodic parameter for each minor phrase using lexical or grammatical features of the phrase. Then, the second set of rules, D-rule, adjusts the prosodic parameter to the dialog context using dialog information.

The total framework of modeling contextual effects of prosody in dialog is depicted in Figure 1. Abe and Mizuno have proposed a similar paradigm of two-step modeling of prosody for changing the speaking style[7]. Our proposed method is characterized by the successive procedure of rule generation through the use of prediction errors. The T-rule is derived from isolated utterances based on a stochastic modeling technique. Then, the T-rule is applied to utterances in dialogs. It captures prosodic characteristics of isolated utterances very well. However, the T-rule knows nothing about how to predict prosody appropriate to

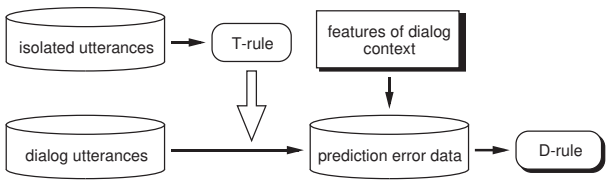

Figure 1. Block diagram of modeling contextual effects

the dialog context. Most errors in predicting prosody of the dialog utterances are caused by the contextual effects of the dialog. The D-rule is derived from these prediction errors based on a stochastic modeling technique again. This two-step modeling has the following advantages.

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

2.2. T-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[6]. They modeled the F0 maximum, called the global model in their paper, using the linear regressive method. We used the same stochastic modeling technique as theirs to obtain the T-rule. Each minor phrase was described in terms of 7 syntactic features. They include modification relationship to the preceding and the following phrases, accent type of the preceding, the current and the following phrases, 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 dialog context using the linear regression method, too. Training cases for D-rule modeling are prepared for each minor phrase in dialog utterances. We investigated characteristics of the dialog context and came up with a set of 7 features reflecting the context. Table 1 lists all the features of dialog information and their values. Because the *DF6* and the *DF7* are features assigned not for each minor phrase but for an utterance, all minor phrases in an utterance have the same value for the *DF6* and the *DF7*. Other features, in general, have different values in an utterance. The training cases are described in terms of a prediction error value and 7 features of dialog information.

Table 1. Features of dialog information

ID	feature description [values]
DF1	contain newly introduced words [new-proper-noun, new-noun, given]
DF2	after filler [yes, no]
DF3	before filler [yes, no]
DF4	function of the phrase [major-point-of-answer, addition-of-information, no]
DF5	detailed part-of-speech [*12 categories*]
DF6	emotional function [surprised, disappointed, no]
DF7	semantic relationship to the preceding utterance [proceeding, elaboration, continuation, answer, acknowledgment, confirmation]

3. EXPERIMENTS

3.1. Speech Data

This method for modeling prosodic parameters in dialog needs two kinds of utterances; isolated utterances which are independent of dialog context and dialog utterances which are contextualized in dialog.

The material for isolated utterances is a set of 503 phonetically balanced sentences designed at ATR. These sentences were 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.

In order to collect contextualized utterances, two kinds of dialogs were carried out under different conditions. The first condition gives simulated dialogs in which the speakers spontaneously utters with a definite goal. One dialog was recorded under this condition with a task of the route inquiry. The utterances in this condition are referred to Dialog-I. In the second condition, the speakers carry out dialogs based on scripts. Although these utterances are not spontaneous, some utterances contained disfluencies, such as fillers, repairs, and so on. The speakers were instructed that they did not completely have to follow the scripts and they were allowed to change the utterance sentences in their own ways keeping the meaning of the utterance. Two dialogs were recorded as the second kind of dialogs, where the task was information service for tourists. The utterances in this script-based dialog are referred to Dialog-II. One participant in all dialogs was the same speaker as that of isolated utterances. Only his utterances were used for the D-rule modeling. The maximum F0 frequencies for minor phrases were also manually identified for the dialog utterances. The number of dialog utterances was 86 and 49 for Dialog-I and -II, respectively.

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, Dialog-I, and -II was 3288, 250, and 220, respectively. The contents of three kinds of utterances are completely different.

3.2. Performance of T-rule

Before discussing total prediction result of the F0 parameters, let us evaluate performance of the T-rule. Table 2 shows prediction errors by the T-rule. The T-rule was derived from the 503 isolated utterances. Table 2 (1) indicates the fundamental performance of the T-rule. The T-rule was evaluated for the isolated utterances based on 10-fold cross validation. The former experiment by Abe[6] is comparable to this result

Table 2. Performance of T-rule

	training data	test data	averaged error [Hz]
(1)	isolated	isolated	10.0
(2-I)	isolated	Dialog-I	20.1
(2-II)	isolated	Dialog-II	17.4
(3-I)	Dialog-I	Dialog-I	18.7
(3-II)	Dialog-II	Dialog-II	14.7

Table 3. Performance of D-rule

(a) Dialog-I

	test data	averaged error [Hz]	
		only T-rule	T- & D-rule
(1)	all	20.1	19.6
(2)	E30*	41.4	37.1

(b) Dialog-II

	test data	averaged error [Hz]	
		only T-rule	T- & D-rule
(1)	all	17.4	16.2
(2)	E30*	38.9	27.2

(*E30: the cases of which the prediction error by the T-rule is larger than 30Hz.)

on a different speaker. The multiple correlation coefficient in our experiment was 0.784. It is a little bit lower than Abe's result 0.843. We consider that this decreased fitness of the model was due to the difference of the speakers. The speaker in Abe's experiment was professional, but not in our experiment. However, the prediction error was smaller than Abe's RMS error 10.0Hz. It suggests that our speaker uttered monotonously.

The prediction error by the T-rule increases to 20.1 Hz and 17.4Hz from 10.0Hz for Dialog-I and -II, respectively, as shown in Table 2 (2). This differences between (1) and (2) were mainly caused by the contextual effects of the dialog. For reference, Table 2 (3) indicates the closed evaluation results of the T-rule sets which were derived from Dialog-I and -II.

3.3. Performance of D-rule

The prediction error data which the T-rule resulted for minor phrases in dialogs 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 dialog information. Dialog-I and -II were separately used for learning the D-rule.

The evaluation was carried out by using 10-fold cross validation. The top rows (1) in Table 3 (a) and (b) show that the D-rule slightly decreased the total prediction error of F0 maxima over all cases of the dialog utterances. From this result, we cannot say that D-rule is very effective. However, all utterances or all minor phrases do not necessarily have contextual effects of the dialog. We can expect that cases which have large prediction errors by the first T-rule application are much influenced by the dialog context. The D-rule was again evaluated only using the cases which have initial prediction errors larger than 30Hz. Fifty seven and thirty three cases were matched with this condition for Dialog-I and -II, respectively. Note that the D-rule learning was carried out using all the cases in each set of the dialog utterances. The bottom rows (2) in Table 3 (a) and (b) show the second evaluation. The D-rule decreased total prediction errors by 4.3Hz and 11.7Hz for Dialog-I and -II, respectively.

Table 4. Partial correlation efficients of the D-rule

feature	partial correlation efficient	
	Dialog-I	Dialog-II
DF1	0.148	0.200
DF2	0.090	0.104
DF3	0.078	0.017
DF4	0.175	0.185
DF5	0.145	0.405
DF6	0.264	0.284
DF7	0.261	0.406

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. If it is 1.0, the feature set completely captures the behavior of the target parameter, that is the F0 maximum of the minor phrase in our problem. The D-rule sets score 0.468 and 0.574 for Dialog-I and -II, respectively. These small coefficients suggest more features of dialog information should be supplemented. Table 4 shows the partial correlation coefficients of each feature. If a partial correlation coefficient is large, its feature gives a lot of contribution to predicting the target parameter. We can find the same tendencies in two D-rule sets from Dialog-I and -II. The *DF7*, the relationship to the preceding utterances, had large partial correlation coefficients and it is very important to predict F0 pattern in dialog. The coefficients of the *DF2* and the *DF3* are very small. The position or the existence of disfluencies has less contextual effects on F0.

Figure 2 depicts details of model parameters of the linear regressive method for the D-rule derived from Dialog-II. The '*surprise*' and the ' *disappointment*' for the *DF6* give large boost of F0. However, this is not reliable because only a few utterances contained these emotional feelings. As for the *DF1*, new nouns decreases F0 while new proper nouns increase F0. All nouns which first appeared through a session of the dialog was labeled by the '*new noun*'. This labeling reduced the effects of the '*new noun*'. Figure 2 (g) shows that the whole answer utterance has low F0. Note that the *DF7* feature assigned to an utterance and it gives the same value to all minor phrases in the utterance. However, F0 is partially boosted at the major points of the answer utterance as shown in (d).

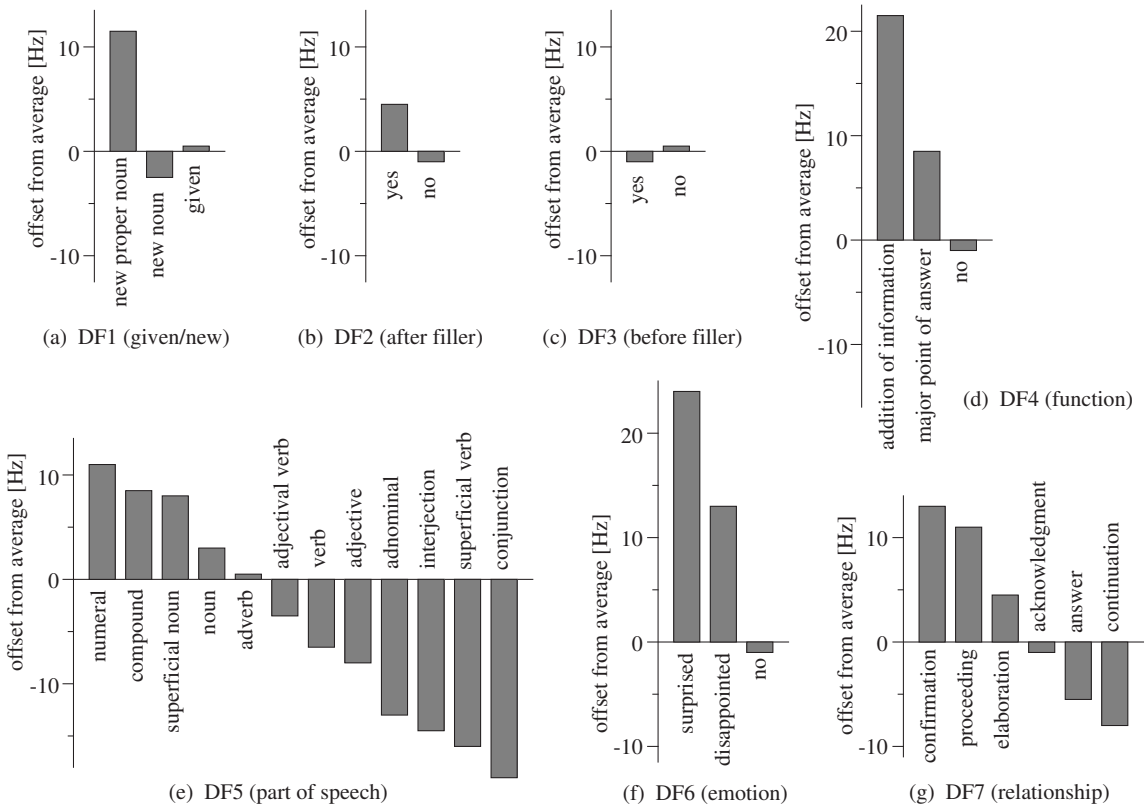

Figure 2. Model parameters of D-rule

4. CONCLUSIONS

This paper proposed a new method of quantitatively modeling the F0 parameter of dialog utterances. The rules which control the F0 parameter with accordance with the dialog information are derived by two-step paradigm of rule generation. The fundamental prosodic characteristics of dialog utterances are removed by conventional rules of TTS before the dialog information is considered. Even if the amount of dialog utterances is not so large, efficient rules was derived. The D-rule decreased the total prediction error by about 10Hz for utterances of which the prediction error by the T-rule was large. However, the multiple correlation coefficient of the D-rule is rather small and the model is not sufficiently precise. It is necessary to add new features of dialog information to the D-rule model.

This paper also shows that the spontaneous dialogs are not good materials for obtaining rules which quantitatively predict prosodic parameters based on the the dialog information. The efficient use of the dialog information is indispensable to speech synthesis of natural utterances in dialog systems. In general, the spontaneous speech contains a lot of fluctuations of the prosodic parameters. It can be expected that good rules are derived form well-organized dialogs, such as script-based dialogs.

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] J.M. Garrido, J. Llisterri, C. de la Mota and A. Rios: "Prosodic Differencies in Reading Style: Isolated vs. Contextualized Sentences", *Proc. of Eurospeech '93*, pp.573-576 (1993).

[4] J. Hirschberg and B. Grosz: "Intonation Features of Local and Global Discourse Structure", *Proc. of DARPA Speech and Natural Language Workshop*, pp.441-446 (1992).

[5] N.A. Daly and V.W. Zue: "Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech", *Proc. of ICSLP '92*, pp.763-766 (1992).

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

[7] M. Abe and H. Mizuno: "A Strategy for Changing Speaking Styles in Text-to-Speech Systems", *Proc. of 2nd ESCA/IEEE Workshop on Speech Synthesis*, New York, pp.41-44 (1994).