On the Global F0 Shape Model using a Transition Network for Japanese Text-to-Speech Systems

Yasushi Ishikawa and Takashi Ebihara

Information Technology R&D Center, MITSUBISHI Electric Corporation
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan
Phone: +81-467-41-2077 FAX: +81-467-41-2136
{yasushi,ebi}@media.isl.melco.co.jp

Abstract

In this paper, we describe a model of fundamental frequency control. In general, a two stage model which consists of a global model and a local model is used as a F0 control method for Japanese text-to-speech systems. We propose a model which is represented by transition network as a global model that generates parameters of a local pitch model from linguistic parameters of a sentence. In the proposed model, syntactic analysis and generation of F0 parameters are integrated, and the nodes of a network represent the linguistic and prosodic state of a sentence. The parameters of a local model is generated when taking transition. We also propose a training method of the network. The prediction results showed our model can predict the phrasal accent parameters with satisfactory high accuracy. We also describe the model can be applied prediction of pause position.

1. INTRODUCTION

A model of fundamental frequency (F0) control is one of the most important problems for the naturalness of synthesized speech in Japanese TTS systems. In general, a two stage model which consists of a global model and a local model is used as Japanese F0 control model [1]. A local model is a model that generates F0 contours of an accent phrase. Fujisaki model [2] is one of the typical models. A global model generates parameters of a local model from linguistic features and other factors which are obtained by linguistic processing of an input sentence.

Recently, global models based on F0 contour prediction using statistical method are proposed and good results were reported [1,3,4]. In these studies, quantitative relation between F0 contours and linguistic parameters of an input sentence from large database. Thus, linguistic processing which analyzes syntactical structure or semantic dependency of phrases quantitatively is required. However it is very difficult to realize robust linguistic analysis, and even if successfully analyzed, it is also difficult to represent linguistic structure of a sentence or contextual features of phrases with quantitative parameters.

Our goal is to realize a prosodic model which can represents the relation between linguistic and prosodic features and is based on a robust linguistic processing. In this paper, a global model that integrates F0 parameter generation with linguistic analysis, and training method of the model are proposed. And we show that a proposed method is also successful for prediction of pause position.

2. NETWORK MODEL FOR GENERATION OF PROSODIC PARAMETERS

2.1 Network Model

The global model is required to predict parameters which represent global F0 shape with high accuracy. In Japanese a global model generates F0 parameters from linguistic features such as semantic dependency of phrases, part of speech, in general. Thus the linguistic processing which extracts such linguistic parameters is required to analyze any sentences including ill-formed sentences robustly. Our basic idea is to relate linguistic structure with prosodic features directly without quantification or classification of linguistic features of a sentence.

The proposed global model is represented by transition network as shown in Figure 1. Each arc is labeled with a phrase category, and each node represents not only syntactical but also prosodic state of a

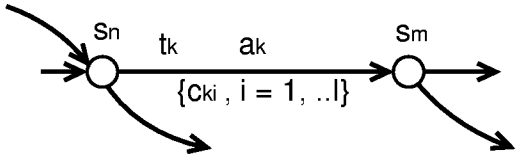

Figure 1. The transition network

sentence. Starting at the beginning of the sentence, each phrase is compared with categories labeled on the arc t_k from the initial state. If the phrase and a category C_{ki} match, a state shifts to the next node S_m, and the model generates F0 shape parameter a_k when taking transition.

2.2 Training Algorithm

The network model is trained by iterating the split of a node and the arc in the network [5]. The training algorithm is shown below.

Step 1. Initial Network

An initial network is created as very simple form. Figure 2 shows an example of the initial network. In this network the initial node **S1** is assigned beginning of a breath phrase and the final node **SN** is assigned end of a breath phrase. All arc is labeled with all phrase categories. The output F0 parameter of an arc is obtained by averaging F0 parameters of all phrases which pass this arc in the learning data.

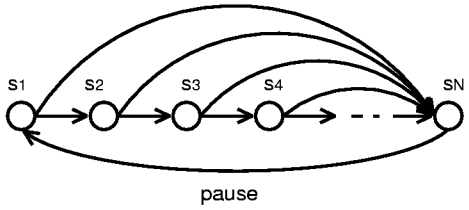

Figure 2. The Initial Network

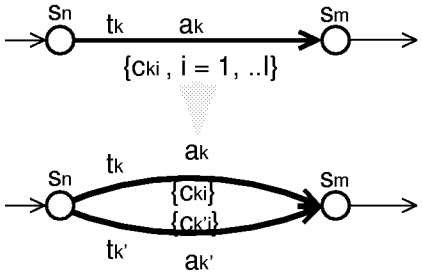

Figure3. Arc splitting (split in the labels)

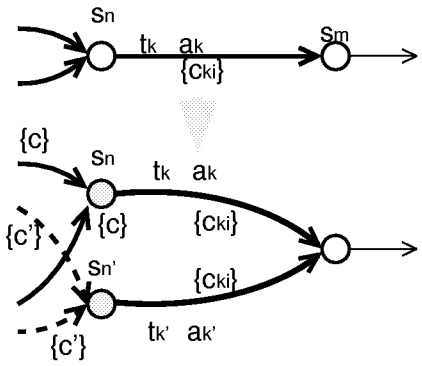

Figure4. Node splitting
(split in the contextual domain)

Step 2. Detection of the arc with maximum error.

Comparing the output value with learning data, the arc with maximum prediction error is detected.

Step 3. Splitting

Two kinds of splitting are applied on the arc with maximum error.

Step 3-A Arc splitting

Classifying categories on the arc into two sets, the arc is split in order to minimize prediction error. This splitting in label domain means to learn the difference of F0 shape that is due to a linguistic category of a phrase (Figure 3).

Step 3-B Node splitting

Classifying categories of the preceding phrases of phrases on the selected arc, the preceding node of the arc is split. This splitting means to learn the difference of F0 shape caused by the phrase context.

Step 4. Select best splitting and retraining

Splitting which minimizes prediction error is selected and output values of modified arcs is obtained by averaging training data..

Iterating step 2 through 4, the network which can represent linguistic and prosodic features of sentence will be obtained.

3. EXPERIMENT

3.1 Experiment Data

The experiment was carried out in order to evaluate the proposed model and training algorithm. In the experiment 503 phonetically balanced sentences uttered by a male professional announcer are used. A F0 contour generation model based on linear approximation is adopted as s local pitch generation model. An accent component parameter shown as **a** in figure 5 is used as a parameter to be predict. The parame-

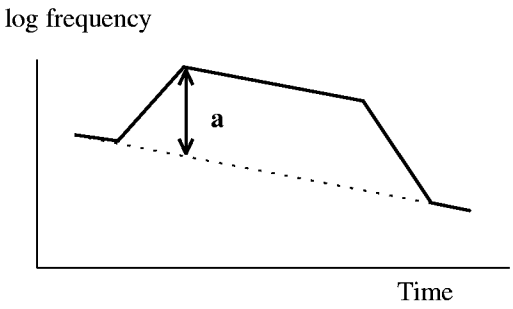

Figure5. The parameter of the local model

ters were detected manually observing pitch pattern. Training was carried out using learning data of 400 sentence utterances (3073 phrases) , and the model were evaluated using the remaining 103 sentences (650 phrases).

3.2 Linguistic category

We compared several sets of linguistic categories in previous work, the results of experiments show that better performance is obtained with the linguistic category based on uses of a part of speech rather than direct category of a part of speech. Table 1 shows linguistic labels which are used in the experiment. And in this experiment breath group boundaries are given.

3.3 Results

Figure 6 shows prediction errors of learning data and test data. In this figure the vertical axis shows prediction error in octave, the horizontal axis shows the number of splitting. The results that the prediction errors decrease with a number of training show efficiency of the proposed method.

This model represent only relation between linguistic and intonation features. However it is well known that there is obvious relation between F0 parameter shown in figure 5 and some other factors. The length of accent phrase is one of the typical factors. Thus, we carried out other experiment in which F0 parameters are modified by length of phrase. At first relation between length and F0 parameter was obtained from learning data using statistical method. Training and evaluation are carried out with F0 parameters which are modified with obtained coefficients shown in table 2. The errors are shown in figure 7.

We also evaluated a conventional statistical method in which nine control factors and 28 categories in total are used. The factors include length of phrase, a phrase boundary type, a part of speech and semantic dependency of phrases. The classification of control factors was obtained assuming the ideal linguistic analyzer. The prediction errors by the statistical method are shown in table 3. The results shows the accuracy of proposed method is better than the conventional statistical method.

Table 1. Linguistic categories of a phrase

category
noun + /no/, /eno/ (pp. possessive case !K
noun + /wa/, /ga/ ... (pp. nominative, objective)
noun + other
declinable word phrase
declinable word phrase (2nd phrase in compound phrase)
others

*pp.: postpositional particle of Japanese

3.4 Prediction of Pause Position

Since the proposed model represents relation between linguistic and prosodic features, it can be applied to prediction of other prosodic parameter. On the other hand, obvious correlation between prosodic parameters, such as pause and pitch. Thus it is considered that the model which trained to represent F0

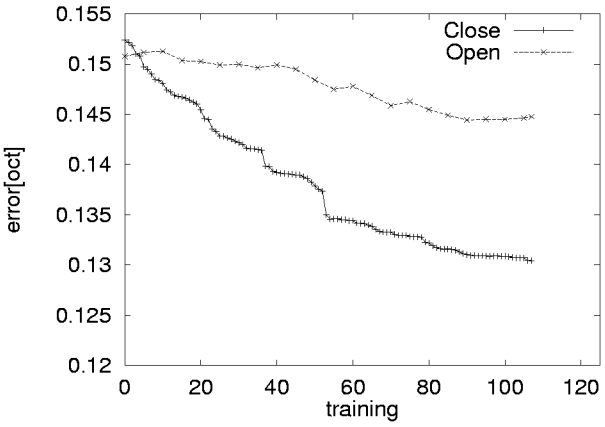

Figure 6. Prediction errors

Table 2. Modification coefficients

length in mora	1,2	3	4	5	6	other
coef. (oct)	0.22	0.04	-0.02	-0.04	-0.07	-0.11

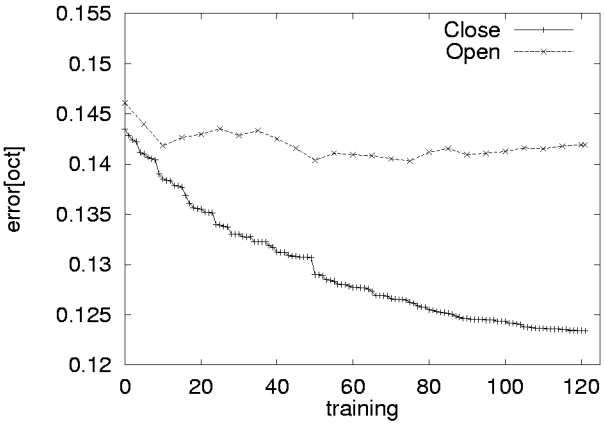

Figure 7. Prediction error after modification

Table 3. Prediction error by the statistical method

Data Set	prediction error [oct]
closed	0.153
open	0.148

features can represent other prosodic features. We carried out the experiments that predicts position of pause in a sentence. In this experiment, an arc of transition network has probability of pause generation shown in figure 8. In the figure p denotes the probability that pause occurs and ($1 - p$) is the probability that pause does not occur. Since it is difficult to decide the optimum positions of pause in a sentence from possible transitions, we adopted simple decision method that a pause is generated at the transition with highest probability in first N phrases. The results are shown in figure 9 and 10. About 75 % of pause position are correctly predicted in both learning and test data set. This results show the efficiency of the proposed model as a model that represents linguistic and prosodic features in Japanese.

4. CONCLUSION

In this paper, the model for global F0 shape based on transition network and the training algorithm of the model were proposed. The proposed model with linguistic parameters which are considered on the uses of a part of speech archived high performance. The accuracy of prediction is higher than the conventional model based on statistical method. And the model can represent not only f0 shapes but also pause position well. It strongly suggests that natural synthesized speech can be obtained with this model. Our model also has the following advantages

- The model requires small amount of computation.
- The model is a powerful representation for grammars. Any sentences includes ill-formed sentences can be analyzed with this method.

Future work includes the following enhancements: (1) allowing recursive arcs for more efficient and accurate model; (2) introducing semantic parameters, (3) evaluation of synthesized speech.

ACKNOWLEDGMENT

The authors would like to thank Mr. Nakajima, the head of a department, for his encouragement and support of this research.

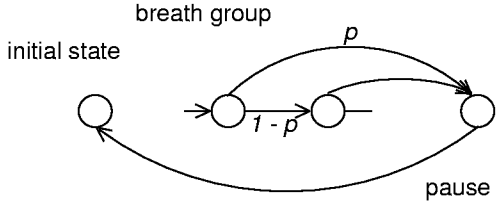

Figure 8. Network with probabilities p that pause occurs

REFERENCES

[1] M. Abe and H. Sato, "Two-Stage F0 Control Model using Syllable based F0 Units," ICASSP'92 pp.II-53 - II-56, 1992

[2] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Jpn(E), 5, pp.233-242 1984

[3] Sagisaka,Y. "On the Prediction of Global F0 shape for Japanese Text-to-Speech", Proc. ICASSP90 pp.235-328, 1990

[4] Traber C. "F0 Generation with a database of Natural F0 Patterns and with a Neural network:, Proc. ESCA Speech Synthesis Workshop pp.141-144, 1990

[5] Takami,J. and Sagayama,S., "A Successive State Splitting Algorithm for Efficient Allophone Modeling", ICAASP'92, pp.573-576 (1992)

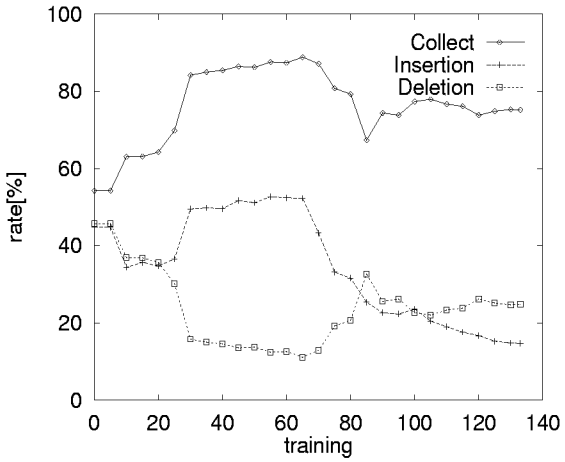

Figure 9. Accuracy of pause prediction (learning data)

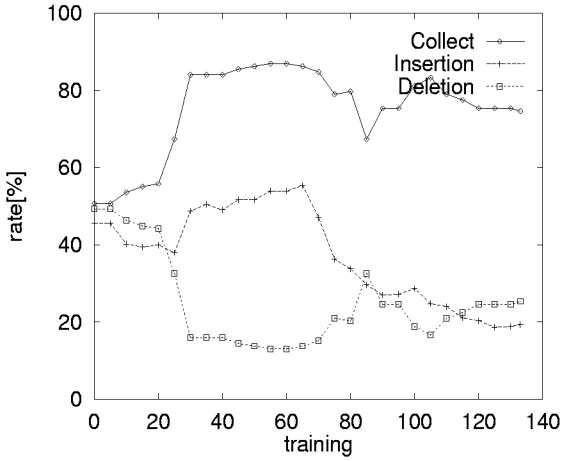

Figure 10. Accuracy of pause prediction (test data)