GENERATION OF F_0 CONTOURS USING A MODEL-CONSTRAINED DATA-DRIVEN METHOD

A. Sakurai

Texas Instruments Japan Ltd.
Tsukuba R&D Center
Miyukigaoka 17, Tsukuba, Japan

K. Hirose, N. Minematsu *

The University of Tokyo
School of Frontier Sciences
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan

ABSTRACT

This paper introduces a novel model-constrained, data-driven method for generating fundamental frequency contours in Japanese text-to-speech synthesis. In the training phase, the parameters of a command-response F_0 contour generation model are learned by a prediction module, which can be a neural network or a set of binary regression trees. The input features consist of linguistic information related to accentual phrases that can be automatically derived from text, such as the position of the accentual phrase in the utterance, number of morae, accent type, and parts-of-speech. In the synthesis phase, the prediction module is used to generate appropriate values of model parameters. The use of the parametric model restricts the degrees of freedom of the problem, facilitating data-driven learning. Experimental results show that the method makes it possible to generate quite natural F_0 contours with a relatively small training database.

1. INTRODUCTION

One of the most important and difficult problems in current text-to-speech (TTS) systems is the generation of natural-sounding F_0 contours from text. Whereas rule-based methods have provided suboptimal but accepted solutions for many years, the increasing availability of fast and low-cost hardware, as well as the advent of databases containing prosodic information (prosodic databases) have paved the way for data-driven approaches to intonation modeling.

However, the construction of prosodic databases is in general labor-expensive due to the great amount of hand work involved in the labeling process, and data-driven approaches proposed so far have not been successful in achieving efficient utilization of training data resources.

In this paper, we propose an efficient data-driven method for F_0 contour modeling and generation based on a parametric command-response model (hereinafter referred to as the F_0 model [1]), whose parameters are predicted by a prediction module. The idea is to utilize a model instead of dealing directly with the F_0 contour in order to bridge the gap between linguistic and prosodic features. The model would create constraints to reduce degrees of freedom and improve learning efficiency.

The F_0 model has been widely used in rule-based TTS systems due to the good correspondence between its parameters and syntactic features, which suggests the existence of a mapping that can be statistically modeled. Another advantage of the model is the small number of parameters required to represent F_0 contours, which potentially improves the efficiency of data-driven learning methods. The use of the F_0 model combined with learning techniques has already been proposed in other works[2][3], but they require parameters that cannot be easily obtained automatically as the input of a TTS system. In the present work, the input parameters contain only information that can be automatically derived from text.

The following sections explain details of the proposed method. Section 2 contains a brief description of the F_0 model. Section 3 describes how F_0 contours are parametrized using the F_0 model, specifying the input and output features of the prediction module. Section 4 contains a description of the prosodic database used. Section 5 describes the two learning methods used – neural networks and binary regression trees –, and shows the results of some evaluation experiments. Finally, informal listening tests comparing generated F_0 contours are described in Section 6, and some comments and additional remarks are given in Section 7.

2. PARAMETRIC REPRESENTATION OF F_0 CONTOURS USING THE F_0 MODEL

The F_0 model [1] is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse called phrase command, and

* The University of Tokyo, School of Engineering

the accent component is generated by another second-order, critically-damped linear filter in response to a step function called accent command. The F_0 model is given by the following equation:

$$\begin{aligned} lnF_0 = \quad & lnF_{min} + \\ & \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \\ & \sum_{j=1}^{J} A_{aj} \{G_{aj}(t - T_{1j}) + \\ & - G_{aj}(t - T_{2j})\} \end{aligned} \quad (1)$$

In the equation above, $G_{pi}(t)$ and $G_{aj}(t)$ represent respectively phrase and accent components. F_{min} is the bias level, I is the number of phrase commands, J is the number of accent commands, A_{pi} is the magnitude of the ith phrase command, A_{aj} is the amplitude of the jth accent command, T_{0i} is the instant of the ith phrase command, T_{1j} is the beginning of the jth accent command, and T_{2j} is the end of the jth accent command. The F_0 model also makes use of other parameters (α_i and β_j) to express G_{pi} and G_{aj}, but in this work they are respectively fixed at 3.0 and 15.0. The bias level F_{min} is fixed at 51.0 Hz for the voice of the speaker used in the experiments.

3. INPUT AND OUTPUT FEATURES

The input features to the prediction module consist of information that can be automatically extracted from text in synthesis time. They are related to accentual phrases, so the module assumes that the text has been previously split into accentual phrases. In a rule-based system, typical features used to determine the value of F_0 model parameters are: position of accentual phrases, number of morae in the accentual phrase, parts-of-speech, time elapsed since the last command, etc. In the present paper, the input features used are listed in Table 1.

In the table, position of the accentual phrase refers to its position in the sentence. Mora is a rhythmic unit in Japanese similar to a syllable, and the accent type classification is based on the position of the mora containing an F_0 downfall (accent nucleus). An accentual phrase is of type 1 when the accent nucleus is located on the first mora, and so on. An accentual phrase that does not contain an accent nucleus is classified as type 0. The term "word" refers to the smallest meaningful unit in Japanese grammar. An accentual phrase is usually made up of one or more words. A word can be classified according to its part-of-speech (POS), and some POS classes admit multiple conjugations (7). We selected as input features the POS and conjugation of the first and last words of the accentual phrase, which can be considered the most relevant.

Table 2 contains the list of output features to be predicted by the prediction module. The indices i and j are eliminated from the notation, since the prediction occurs individually for each accent command, which may be preceded or not by a phrase command.

Table 1. Accentual phrase features used as inputs to the prediction module

Input feature	Maximum value
Position of accentual phrase	18
No. of morae	15
Accent type	9
No. of words	8
POS of first word	21
Conjugation of of first word	7
POS of last word	21
Conjugation of last word	7

Table 2. Output features of the prediction module

Output feature	Symbol
Phrase command magnitude	A_p
Accent command amplitude	A_a
Offset of T_0	t_{0off}
Offset of T_1	t_{1off}
Offset of T_2	t_{2off}
Phrase command flag	PF

In the table, $t_{0_{off}}$ is the offset of T_0 with respect to the segmental beginning of the accentual phrase. $t_{1_{off}}$ and $t_{2_{off}}$ are respectively offsets of T_1 and T_2 with respect to segmental anchor points. These anchor points are respectively defined as the beginning of the first high mora for t_{1off}, and the end of the mora containing the accent nucleus for $t_{2_{off}}$. The first high mora of the accentual phrase is either the first mora for accentual phrases of type 1, or the second mora for accentual phrases of other types. The phrase command flag (PF) is a binary flag that signals the occurrence of a phrase command at the beginning of the accentual phrase.

4. THE PROSODIC DATABASE

The prosodic database is made up of 486 sentences extracted from ATR's continuous speech database [4] (speaker MHT). The database has been divided into three parts: 388 sentences constitute the training section, 50 sentences constitute the validation section, and 48 sentences constitute the

Table 3. Mean square error of F_0 contours generated by neural network with respect to natural speech

Neural net configuration	No. of elements in hidden layer	MSE
MLP	10	0.218
MLP	20	0.217
Jordan	10	0.220
Jordan	20	0.215
Elman	10	0.214
Elman	20	0.232

test section. The validation section is used in the neural network training as a measure to avoid overtraining. The timing parameters of the training data are obtained as suggested in [5], and then an analysis-by-synthesis process using F_0 contours extracted from natural speech is carried out on the magnitudes of phrase commands and amplitudes of accent commands.

5. PREDICTION MODULE

5.1. Prediction Module Based on a Neural Network

Neural networks provide a good solution for problems involving strong non-linearity between input and output parameters, and also when the quantitative mechanism of the mapping is not well understood.

The use of neural networks in prosodic modeling has been reported in [6] and [7], but those methods do not make use of a model to limit the degrees of freedom of the problem. Another difference is that they are syllable-based, and additional care must be taken in order to account for the continuity of F_0 contours (using recurrent networks). In our modeling, the continuity and basic shape of F_0 contours are ensured by the F_0 model.

In this work, three types of neural network structures are evaluated: the multi-layer perceptron (MLP), Jordan (a structure having feedbacks from output elements), and Elman (a structure having feedbacks from hidden elements). The latter two neural network structures are called partial recurrent networks, and are tested here in order to account for the mutual influence of neighboring accentual phrases. All structures have a single hidden layer containing either 10 or 20 elements.

For the experiments, we utilized the SNNS neural network simulation software [8]. The results of F_0 contour prediction on the test data set are shown in Table 3. In the table, the MSE (mean square error) value corresponds to the average squared difference between the generated F_0 contour and the contour extracted from natural speech, in log scale.

Table 4. Evaluation of F_0 contours generated by regression trees

Stop criterion	MSE
10	0.218
20	0.222
30	0.210
40	0.217
50	0.220

5.2. Prediction Module Based on Binary Regression Trees

A disadvantage of neural networks is the difficult interpretability of the resulting prediction module. In order to obtain a basis for comparison and also to have an idea of the individual influence of each input feature in future works, we also solve the same F_0 model parameter prediction problem using binary trees, and carry out a comparison between the results in terms of MSE measure. The human-interpretable results provided by binary regression trees can also be eventually reflected in the design of neural networks.

For the experiments, we use the freeware Wagon [9] from the Edinburgh Speech Tools Library. Trees are constructed using different values for the stop criterion, which is the minimum number of examples per leaf node (10, 20, 30, 40, 50). Note that one regression tree is needed for each parameter to be predicted.

The results of F_0 contour prediction using tree regression are shown in Table 4. In terms of MSE measure, we note that there is no significant difference with respect to the results obtained with the neural network prediction module.

An example of F_0 contour generated by a set of regression trees with stop criterion set to 30 is shown in Figure 1. In the figure, the first frame shows a speech waveform corresponding to the utterance, the second frame shows phoneme labels, the third frame shows the generated F_0 contour (continuous line) and the reference contour extracted from natural speech (discontinuous line), and the fourth frame shows the predicted F_0 model parameters.

6. LISTENING TESTS

We carried out listening tests using natural speech samples with modified F_0 contours. F_0 contour modification was done using an LMA filter[10]. Speech samples corresponding to 9 sentences were taken, and their F_0 contours were modified by applying two different F_0 contours: an F_0 contour generated by an Elman network containing 10 elements in the hidden layer, and another one generated by a set of regression trees with stop criterion set to 30. These are the configurations that yielded the best scores in terms of MSE distance with respect to natural speech. The 9 resulting pairs

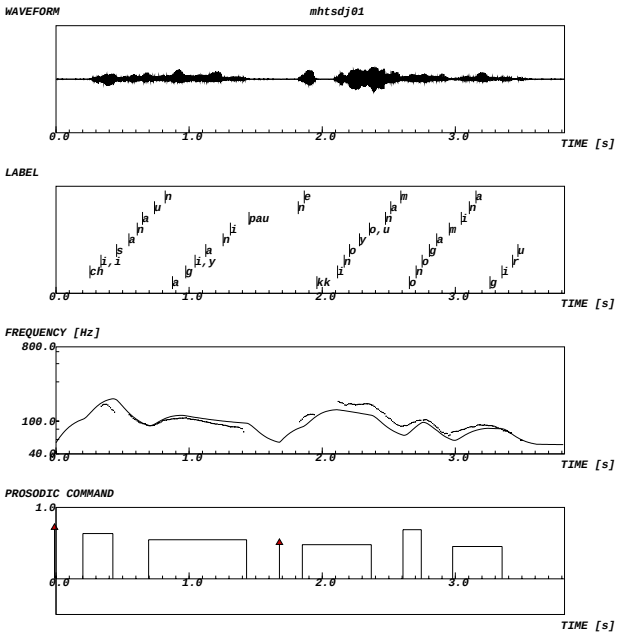

Fig. 1. Example result using a set of regression trees

of speech samples resynthesized with both prediction modules were presented to 9 listeners, who were asked to select the one that sounded better. From the total of 81 speech pairs presented, in 30 pairs the preferred F_0 contour was the one generated by the neural network, against 19 of the set of binary trees. The listeners could not tell their preference in the remaining 32 pairs. The results show that the listeners could not detect a significant perceptual difference between the two prediction modules, and the general impression was positive for both.

7. COMMENTS

We presented a modeling and generation scheme of F_0 contours for Japanese TTS that uses a superpositional command-response model and a prediction module, which can be a neural network or a set of binary regression trees. The model reduces the number of degrees of freedom and facilitates data-driven learning. Compared to rule-based methods, the approach enables easier construction of a high-quality F_0 contour predictor, reducing dependence on ad-hoc rules. In addition, compared to other statistical approaches that do not make use of an F_0 contour generation model, the encoded representation of F_0 contours enables the utilization of more compact prosodic databases for training. For now on, further investigation is necessary on the contribution of different input parameters, and also on other possible learning methods and configurations for the prediction module.

8. REFERENCES

[1] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *J.Acoust.Soc.Jpn(E)*, vol. 5, no. 4, pp. 233–242, 10 1984.

[2] T. Hirai, N. Iwahashi, N. Higuchi, and Y. Sagisaka, "Automatic extraction of f0 control rules using statistical analysis," in *Advances in Speech Synthesis*. 1996, pp. 333–346, Springer.

[3] O. Jokisch, H. Mixdorff, H. Krusche, and U. Kordon, "Learning the parameters of quantitative prosody models," in *Proceedings of ICSLP'2000*, 2000.

[4] K. Takeda, N. Sagisaka, S. Katakiri, Abe, and Kurihara, *Speech Database for Research - User's Manual*, ATR, 1988.

[5] T. Hirai and Y. Higuchi, "Automatic extraction of the fujisaki model parameters using the labels of japanese tone and break indices (j_tobi) system," *IEICE Journal D-II (Jap.)*, vol. J81-D-II, no. 6, pp. 1058–1064, 6 1996.

[6] C. Traber, "F0 generation with a data base of natural f0 patterns and with a neural network," in *Talking Machines: Theories, Models, and Designs*. 1992, pp. 287–304, Elsevier.

[7] S.H. Chen, S.H. Hwang, and Y.R. Wang, "An rnn-based prosodic information synthesizer for mandarin text-to-speech," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 226–239, 5 1998.

[8] University of Stuttgart, *Stuttgart Neural Network Simulator - User Manual - Version 4.1 - Report no. 6/95*.

[9] P. Taylor and A. Black, *Edinburgh Speech Tools Library - Wagon*, Edinburgh University, http://www.shlrc.mq.edu.au/festival/, 1999.

[10] S. Imai, "Low bit rate cepstral vocoder using the log magnitude approximation filter," in *Proc. of ICASSP'78*, 4 1978, pp. 441–444.