DATA-DRIVEN INTONATION MODELING USING A NEURAL NETWORK AND A COMMAND RESPONSE MODEL

* †*Atsuhiro Sakurai,* †*Nobuaki Minematsu, and* †*Keikichi Hirose*

*Tsukuba R&D Center, Texas Instruments Japan
7 Miyukigaoka, Tsukuba, Ibaraki, 305-0841, Japan
†Dept. of Information and Communication Engineering, Univ. of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

An intonation modeling scheme for Japanese text-to-speech synthesis is proposed using a command response F_0 model and a neural network to generate F_0 contours of accentual phrases uttered in continuous speech. The neural network is used to predict the values of F_0 model parameters for a whole sentence, focusing on accentual phrases. The features used as inputs to the neural network are: position of the accentual phrase within the sentence, number of morae in the accentual phrase, accent type of the accentual phrase, number of words in the accentual phrase, and parts-of-speech of the first and last words of the accentual phrase. The predicted parameters are: a flag that indicates the presence of a phrase command at the beginning of the accentual phrase, magnitude of the phrase command (if present), amplitude of the accent command, and offset values for the timing of phrase and accent commands. All features are simultaneously predicted. Three types of neural network structures are used, each one with 3 different numbers of elements in the single hidden layer: MLP (multi-layer perceptron), Elman, and Jordan. The method permits efficient prediction of F_0 model parameters, as observed in evaluation experiments and informal listening tests.

1. INTRODUCTION

One of the most difficult problems in text-to-speech (TTS) synthesis is the correct control of prosodic features, especially F_0 contour and duration.

F_0 contours can be generated using superpositional models, which provide accurate approximation of F_0 contours when parameter values are correctly assigned. A representative superpositional model for Japanese F_0 contours has been proposed in [1] as a command-response model (F_0 Model for now on). The model represents F_0 contours in logarithmic scale as the superposition of phrase and accent components, which are associated to different levels in the prosodic structure. This model is used in the present framework as a straightforward way to represent prosodic features in speech databases, as proposed in [2].

However, automatic assignment of F_0 model parameters in real TTS systems is a difficult task due to the complex non-linear dependence on an array of linguistic features. As a consequence, most practical TTS systems have dealt with the problem in a suboptimal way by including a hand-tuned, knowledge-based expert system in charge of predicting F_0 model parameters from text.

It is, though, currently impossible to design an expert system that perfectly mimics a human's intricate F_0 contour generation process. In view of that, considerable research effort has been devoted to the use of statistical techniques to replace human expertise. This is basically the philosophy behind data-driven approaches to intonation modeling. One such attempt is reported in [3]. The method is based on MSR (multiple-split regression) trees, a variation of binary regression tree that encompasses both binary split regression and multivariate linear regression. The method permits the prediction of phrase command magnitudes (A_p) and accent command amplitudes (A_a) when all timing parameters are externally assigned, making use of a measure of deepness of prosodic boundaries.

In this work, we propose a method based on neural networks that yields all F_0 model parameters necessary for the calculation of the F_0 contour, including timing parameter values. The text must be previously split into accentual phrases and undergo morphological analysis, as commonly done by most current TTS systems.

The use of neural networks in prosodic modeling has been reported in [4] (Mandarin) and [5] (German), but they do not make use of a superpositional model. Our idea is that a superpositional model is a good way to reduce degrees of freedom of the system.

Three types of neural network configurations are tested: Elman (a partial recurrent network with feedbacks from hidden units), Jordan (a partial recurrent network with feedbacks from output units), and multi-layer perceptrons (MLP). The following sections describe the method in detail and show the results of evaluation experiments.

2. NEURAL NETWORK MODELING OF F_0 MODEL PARAMETERS

2.1. Network Configuration

The ideal configuration, number of layers, and number of units per layer for a given problem cannot in general be easily pre-established, and often a great deal of heuristics is required to obtain the network configuration [6]. For the present problem, based on previous research, we use two types of partial recurrent networks, commonly known as Elman and Jordan networks, and also the well-known multi-layer perceptron (MLP) (**Figure 1**).

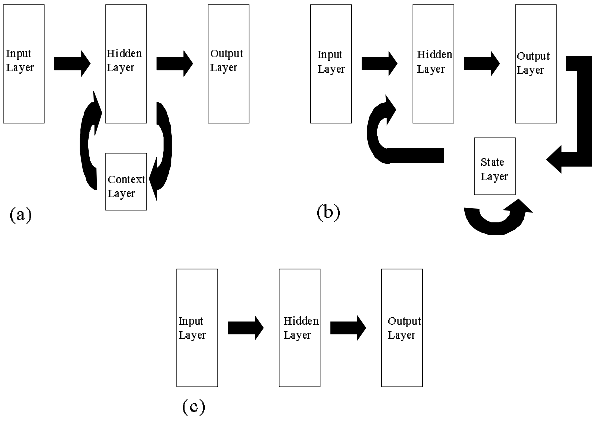

Figure 1: Neural network structures used in F_0 model parameter prediction experiments: (a) Elman network, (b) Jordan network, and (c) multi-layer perceptron (MLP).

2.2. Input and Output Features

The neural network takes as inputs various types of linguistic information that can be directly extracted from text. In a rule-based system, typical features that are used to determine the values of F_0 model parameters are: the position of accentual phrases within the utterance, number of morae, duration, parts-of-speech, time (or number of morae) elapsed since the last phrase command, time (or number of morae) until the end of the sentence, etc. [3,7]. For the present neural network modeling, the features used as inputs to the neural network are listed in **Table 1**, and the output features are shown in **Table 2**. The part-of-speech (POS) classification is based on the ATR Continuous Speech Database [8], and the phrase command flag in **Table 2** is a binary flag that signals the presence of a phrase command before the accentual phrase in question.

Input Feature	Number of classes
Position of accentual phrase within utterance	18
No. of morae in accentual phrase	15
Accent type of accentual phrase	9
No. of words in accentual phrase	8
POS of first word in accentual phrase	37
Conjugation type of first word in accentual phrase	7
Conjugation category of first word in accentual phrase	7
POS of last word in accentual phrase	37
Conjugation type of last word in accentual phrase	7
Conjugation category of last word in accentual phrase	7

Table 1: Neural network input features

Output Feature	Type
Phrase command magnitude (A_p), if exists	Continuous value
Accent command amplitude (A_a)	Continuous value
Phrase command offset ($t_{0\ off}$) , if exists	Continuous value
Offset of accent command onset ($t_{1\ off}$)	Continuous value
Offset of accent command reset ($t_{2\ off}$)	Continuous value
Phrase command flag	Binary

Table 2: Neural network output features

It is worth noting that partial recurrent networks (Jordan and Elman structures) have feedback loops that can model contextual effects such as the dependence of a phrase command on the time (or number of morae) elapsed since the last phrase command, or the interaction between amplitudes of neighboring accent commands, and should theoretically perform better than MLPs. The experiments will show that to some extent, this supposition is true.

3. EVALUATION EXPERIMENTS

3.1. Training on a Prosodic Database

The prosodic database used for training and testing has been designed as proposed in [2]. It is derived from the ATR Continuous Speech Database (speaker MHT) and contains values of F_0 Model parameters which have been automatically calculated using the method proposed in [9] and then slightly hand-corrected.

The database is divided into three sections, respectively used for training, validation, and testing. The training section contains 2803 accentual phrases distributed across 388 sentences, the validation section contains 317 accentual phrases distributed across 50 sentences, and the testing section contains 48 sentences containing 262 accentual phrases. The data effectively used for training comes from the training section. The validation section data are used during the training to provide an independent error measure that indicates when to stop training. All evaluation tests are carried out using the test data section.

Training is carried out using a special backpropagation method for partial recurrent neural networks, and standard backpropagation for MLP. SNNS version 4.1 [10] is used for neural network design and simulation.

3.2. Predicting the Existence of Phrase Commands and Their Parameter Values

Although the existence of phrase commands is signaled by a binary flag, the output of the neural network is a continuous value. For this reason, a threshold is necessary to obtain the phrase flag from the neural network output. The threshold is automatically obtained by increasing its value in steps of +0.01, starting from 0.0, and monitoring the number of insertions and deletions thus obtained. The threshold that approximately balances the number of insertions and deletions is selected.

Table 3 summarizes the results of phrase command prediction obtained with different network configurations. Note that the test data set contains 262 accentual phrases, out of which 111 are accompanied by a phrase command.

Neural network type	Number of elements in hidden layer	Detected (Dt)	Deletion (Dl)	Insertion (In)	Dt/In
MLP	10	83	28	36	2.31
MLP	20	81	30	40	2.03
MLP	50	80	31	34	2.35
Jordan	10	81	30	37	2.19
Jordan	20	79	32	38	2.08
Jordan	50	81	30	36	2.25
Elman	10	81	30	37	2.19
Elman	20	82	29	37	2.22
Elman	50	78	33	37	2.11

Table 3: Phrase command prediction results.

The neural network also predicts the values of the phrase command magnitude A_p and the phrase command offset $t_{0\ off}$, i.e., the time from the phrase command onset t_0 until the beginning of the accentual phrase at the phonetic level. **Table 4** gives the mean-square error for these values with respect to nominal values contained in the database, calculated only over accentual phrases where phrase commands effectively exist.

Neural network type	Number of elements in hidden layer	MSE for A_p (x 10^{-3})	MSE for $t_{0\ off}$ (x 10^{-3} s^2)
MLP	10	30	33
MLP	20	30	32
MLP	50	30	33
Jordan	10	31	34
Jordan	20	30	32
Jordan	50	31	33
Elman	10	31	32
Elman	20	29	32
Elman	50	29	33

Table 4: Results of phrase command parameter prediction

Here, the MSE error is calculated according to equation (1), where p is the nominal value of the parameter and p' is the predicted value:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(p_i - p_i^{'})^2 \qquad (1)$$

3.3 Predicting Accent Command Parameters

Accent command parameters are also predicted by the neural network. **Table 5** gives the MSE for the accent command amplitude A_a, and the offset times for the accent command onset ($t_{1\ off}$) and reset ($t_{2\ off}$). The MSE is also calculated using Equation (1).

Neural network type	Number of elements in hidden layer	MSE for A_a (x 10^{-3})	MSE for $t_{1\ off}$ (x 10^{-3} s^2)	MSE for $t_{2\ off}$ (x 10^{-3} s^2)
MLP	10	29	4.5	4.8
MLP	20	27	5.0	5.3
MLP	50	28	4.9	4.7
Jordan	10	28	4.7	5.1
Jordan	20	25	4.5	4.7
Jordan	50	28	4.2	5.1
Elman	10	28	4.8	4.7
Elman	20	28	4.7	4.8
Elman	50	28	4.4	4.6

Table 5: Results of accent command parameter prediction

3.4 Comparison with F_0 Contours Extracted from Natural Speech

In order to evaluate the overall prediction results of F_0 Model parameters, we obtained the complete set of F_0 Model parameters for the 48 sentences of the test set and compared the F_0 contours produced by those parameters with the F_0 contour extracted from natural speech for each neural network structure.

The results are given in **Table 6**. Note that timing parameters were calculated based on the offsets predicted by the neural network applied to reference points found in the natural speech at the phonetic level. These reference points are: beginning of the accentual phrase (for phrase commands and onsets of accent commands associated to accentual phrases of type 1), beginning of the second mora (for onsets of accent commands associated to accentual phrases of other accent types), end of the mora containing the accent nucleus (for resets of accent commands associated to accentual phrases that contain an accent nucleus), and end of the accentual phrase (for resets of accent commands associated to accentual phrases that do not contain an accent nucleus - type 0). The phrase command reset is always placed at the end of the sentence, and the values of the natural angular frequencies of the phrase and accent control mechanisms (α and β) are respectively fixed to 3.0 and 15.0. The minimum square error between the predicted F_0

contour and the extracted F_0 contour is calculated using Equation (2), where F_0 is the value extracted from the natural speech, and F_0' is the predicted value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}[\log(F_{0i}) - \log(F_{0i}^{'})]^2 \qquad (2)$$

3.5 Informal Listening Tests

In order to further evaluate the results obtained with the method, we applied the F_0 contours generated in the previous item to natural speech samples of the database, using an LMA filter [11]. The generated speech samples show a high level of naturalness, which encourages the implementation of the present method in an actual TTS system.

Neural network type	**Number of elements in hidden layer**	**F_0 Contour MSE $(\log(Hz))^2$**
MLP	10	0.219
MLP	20	0.224
MLP	50	0.225
Jordan	10	0.214
Jordan	20	0.213
Jordan	50	0.226
Elman	10	0.214
Elman	20	0.211
Elman	50	0.232

Table 6: F_0 Model parameter prediction error (MSE)

4. CONCLUSION

We presented an intonation modeling and prediction scheme for Japanese text-to-speech synthesis that predicts the values of F_0 Model parameters using a neural network, and the results show the validity of the method. It was shown that the neural network structures selected in the present experiments are able to predict the whole set of F_0 model parameters, including timing-related values, but the small number of neural network types tested are not sufficient to determine the optimal neural network structure. For now on, other neural network structures should be investigated, and also the effect of scarce training data. In addition, the method should be connected to a TTS system and listening tests must be realized in order to compare it with traditional rule-based systems.

5. REFERENCES

1. H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," *The Journal of the Acoustical Society of Japan (E)*, Vol. 5, No. 4, pp. 233-242 (1984).

2. A. Sakurai, T. Natsume, and K. Hirose, "A linguistic and prosodic database for data-driven Japanese TTS systems," *Proc. of ICSLP-98*, pp. 2843-2846 (1998).

3. T. Hirai, N. Iwahashi, N. Higuchi, and Y. Sagisaka, "Automatic extraction of F_0 control rules using statistical analysis," in *Advances of Speech Synthesis*, Springer, pp. 333-346 (1996).

4. S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 3, pp. 226-239 (1998).

5. C. Traber, "F_0 generation with a data base of natural F_0 patterns and with a neural network," in *Talking Machines: Theories, Models, and Designs*, G. Bailey, C. Benoit, and T. R. Wawallis, Eds. Amsterdam, The Netherlands: Elsevier, pp. 287-304 (1992).

6. R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP magazine*, pp. 4-22 (1987).

7. K. Hirose and H. Fujisaki, "A system for the synthesis of high-quality speech from texts on general weather conditions," *IEICE Trans. Fundamentals*, Vol. E76-A, No. 11, pp. 1971-1980 (1993).

8. K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, Speech Database User's Manual, *ATR Technical Report* (1988-5).

9. T. Hirai and N. Higuchi, "Automatic extraction of the Fujisaki model parameters using the labels of Japanese Tone and Break Indices (J-ToBI) system," *IEICE Journal*, D-II, Vol. J81-D-II, no. 6, pp.1058-1064 (1998) (in Japanese).

10. Stuttgart Neural Network Simulator, User Manual, Version 4.1, Report no. 6/95.

11. S. Imai, "Low bit rate cepstral vocoder using the log magnitude approximation filter," *Proc. of ICASSP'78*, pp. 441-444 (1978-4).