Data-Driven Synthesis of Fundamental Frequency Contours for TTS Systems Based on a Generation Process Model

Keikichi Hirose, Nobuaki Minematsu**, and Masaya Eto**

*Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
** Dept. of Inf. and Commu. Engg, School of Inf. Science and Tech., University of Tokyo
{hirose, mine, eto}@gavo.t.u-tokyo.ac.jp

Abstract

A data-driven method of fundamental frequency (F_0) contour synthesis was developed for Japanese text-to-speech (TTS) conversion systems. In the method, synthesis is done using the F_0 contour generation process model, and the model parameters for each accent phrase are estimated using statistical methods. Although it was already shown that the synthesized F_0 contours sounded highly natural as those using heuristic rules arranged by experts, occasional low quality happened depending on sentences to be synthesized. In the current paper, information on sentence structure, automatically obtainable through the parsing process, is added to input parameters of the statistical methods to obtain a better estimation. The experimental results showed that the new parameter was effective for improving especially phrase component estimation. Furthermore, data-driven estimation of accent phrase boundaries for input text, a necessary step to realize TTS conversion, was also realized in a similar way. The rate of correct estimation reached 90 %.

1. Introduction

Quality of synthetic speech from text-to-speech (TTS) conversion systems is largely improved through recent advancement in speech technology. The improvement, however, is mostly on the segmental aspect of speech, and it rather focused low quality of prosodic features. In view of the success of data-driven methods in speech processing areas, a rather large number of works have been done to generate prosodic features from linguistic inputs using statistical methods, such as neural networks, binary decision trees and so on. When synthesis rules arranged carefully by experts are used, the resulting synthetic speech can have rather high quality, which is hard to be surpassed by statistical methods. This is especially true for fundamental frequency (F_0) contours, for which several models capable of closely approximating natural F_0 movements are already developed. However, to develop synthesis rules for a new style of utterance is time consuming and impossible if the expert has not so much knowledge on the style.

In data-driven methods for F_0 contour generation, F_0 movements can be directly related to linguistic information of the input texts. An HMM-based method succeeded to generate synthetic speech with highly natural prosodic features by counting F_0 delta features [1]. These methods without F_0 model constraints theoretically can generate any type of F_0 contours, but have possibility of causing un-naturalness especially when the training data are limited. Several methods are reported under the ToBI labeling strategy. Constraints by the ToBI system are beneficial in avoiding unlikely F_0 contour being generated. The major problem of ToBI system is that it is not a full quantitative description of F_0 contours, which causes some limitations to the quality of synthesized F_0 contours.

The F_0 contour generation process model (Fujisaki model, henceforth F_0 model) in sentence level [2] will be a good answer for the above problems. It assumes two types of commands, phrase and accent commands, as model inputs, and these commands are proved to have a good correspondence with linguistic (and para-/non-linguistic) information of speech. Advantage of the F_0 model has already been shown in rule-based F_0 contour synthesis in our work on Japanese TTS conversion [3], and other works on several major languages. From this point of view, we adopted F_0 model as constraints in the data-driven synthesis of F_0 contours and obtained good results [4]. Although current constraints are limited to the model's command response features, further constrains are possible based on various knowledge on model commands, such as on command timing as compared to the segmental boundary locations.

The use of F_0 model in a statistical approach was already tried in [5], where multiple split regression trees are used to derive rules to generate F_0 model parameters. However, the timing parameters are excluded from the mapping and have to be externally assigned. Moreover, it uses high-level syntactic information as the statistical model input, which is difficult to be automatically obtained in a TTS system. Our method estimates magnitudes/amplitudes and timings of F_0 model commands from linguistic information automatically obtainable through input text analysis. In our previously reported methods [4], only part-of-speech information was utilized as input parameters representing structure of sentence. In this paper, we newly added "bunsetsu" boundary depth automatically obtained by syntactic analysis using a Japanese sentence parser.

The developed method is based on estimating F_0 model parameters in accent phrase units. In the current paper, we also estimate accent phrase boundaries for input sentences in a framework similar to the F_0 model parameter estimation.
In the following sections, the accent phrase boundary estimation and the F_0 model parameter estimation are explained in section 4 and section 6, respectively.

2. F_0 contour generation from text

In our method of F_0 contour generation, estimation of F_0 model parameters are done for each accent phrase and a sentence F_0 contour is generated using the F_0 model after the estimation process is finished for all the constituting accent phrases. Therefore, given a text, the following 4 processes are necessary before converting it into speech: morpheme analysis, accent phrase boundary detection, accent type estimation of

accent phrases, and F_0 model command estimation. In Japanese, a content word (or words) is concatenated with its following function word(s) to form a "bunsetsu," whose accent type is usually not given in the word lexicon. A "prosodic word" mostly coincides with an accent phrase. We are planning to use free software for the first process, and to adopt a rule-based method for the third process. This paper covers the second and fourth processes, where statistical methods are used.

3. Statistical methods

Three types of statistical methods were used and their performances were compared for the estimation of accent phrase boundaries and F_0 model parameters;

Neural network (NN): Besides the conventional three-layered perceptron (MLP), Jordan (a structure having feedbacks from output elements), and Elman (a structure having feedbacks from hidden elements) networks are also selected to check if the feedback process may have some effects on the prediction accuracy. All structures have a single hidden layer containing either 10 or 20 elements. For the experiments, we utilized the SNNS neural network simulation software [6].

Binary decision tree (BDT): This method has an advantage over neural network methods in that it provides human-interpretable results, which are useful to improve the estimation performance. The freeware Wagon [7] from the Edinburgh Speech Tools Library is used to construct the trees. Stop threshold, represented by the minimum number of examples per a leaf node, is set around 50 according to the result of preliminary experiments.

Multiple linear regression analysis (MLRA): This method is included for the experiments also to obtain human interpretable results. Correlation coefficients can be utilized as indices of input parameter appropriateness.

4. Estimation of accent phrase boundaries

The predictor examines each morpheme boundary of input text, and outputs a binary flag indicating whether the current morpheme boundary is an accent phrase boundary or not. First, an experiment was conducted by selecting part-of-speech, conjugation type, and conjugation form of the morpheme following to the boundary in question as input parameters to the predictor (methods (a)). 503 sentences of the ATR continuous speech corpus (see section 6) were used. Morpheme boundaries and the input parameters used in the experiments were those obtainable from the corpus. The sentences were divided into 3 groups and were used for training, testing and neural network validation process: 388 sentences (6029 morphemes) for training, 48 sentences (543 morphemes) for testing and 50 sentences for validation (not used for BDT and MLRA). Although correct estimation rate was exceeded 88.5 % for all the methods, insertion errors were often observed for compound words as follows:

Correct: | ki gyo ki bo be tsu no | chi N gi N ka ku sa mo | su ko shi zu tsu | chi ji ma Qte |ki ta |

Estimation: | ki gyo | ki bo be tsu no | chi N gi N | ka ku sa mo | su ko shi zu tsu | chi ji ma Qte ki ta |

Here, the sentence can be translated as "Wage differentials according to the size of companies have gradually decreased." Symbol "|" indicates accent phrase boundaries and boundaries before "ki bo betsuno" and "ka ku sa mo" are insertion errors.

In order to cope with these insertion errors, features of the preceding morpheme to the boundary in question were also taken into account (methods (b)). The input and output parameters are summarized in Table1. The experimental results showed some insertion errors including the above two errors can be recovered. The correct estimation rate exceeded 90 % for several statistical methods as indicated in Table 2.

Table 1: Input and output parameters for accent phrase boundary estimation and their numbers of categories for method (b). For method (a), information on the following morpheme is excluded.

Morpheme Features		Categories
Input Parameters	Part-of-Speech of the Preceding Morpheme	21
	Conjugation Type of the Preceding Morpheme	8
	Conjugation Form of the Preceding Morpheme	8
	Part-of-Speech of the Following Morpheme	21
	Conjugation Type of the Following Morpheme	8
	Conjugation Form of the Following Morpheme	8
Output Parameter	Boundary Flag	2 (1 or 0)

Table 2: Insertion and deletion error rates (%) and correct estimation rates (%) of accent phrase boundary estimation (methods (b)). For neural networks, "-10" and "-20" mean the number of elements of the hidden layer being 10 and 20, respectively. "-50" for binary decision tree means the stop threshold being 50.

Methods		Closed			Open		
		Ins.	Del.	Cor.	Ins.	Del.	Cor.
NN	MLP-10	3.0	3.9	93.2	4.5	5.7	89.9
	MLP-20	3.2	3.7	93.1	4.5	5.7	89.9
	Jordan-10	3.2	3.7	93.1	4.5	5.7	89.9
	Jordan-20	3.4	3.5	93.1	4.3	5.1	90.7
	Elman-10	3.0	3.9	93.1	4.5	5.5	90.1
	Elman-20	3.2	3.7	93.1	4.3	5.7	90.1
BDT-50		4.1	4.3	91.6	4.4	5.5	90.0
MLRA		6.3	2.5	91.6	6.3	3.9	89.9

We should note that the errors are not all serious for the F_0 synthesis. For instance, the deletion error before "ki ta" even results in more natural F_0 contours.

5. F_0 model and parametric representation of F_0 contours

The F_0 model is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components [2]. The phrase component is generated by a second-order, critically-damped linear filter in

response to an impulse called phrase command, and 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:

$$\ln F_0 = \ln F_{0\min} + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\} \quad (1)$$

In the equation above, $G_{pi}(t)$ and $G_{aj}(t)$ represent phrase and accent components, respectively. F_{0min} 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 time of the ith phrase command, T_{1j} is the onset time of the jth accent command, and T_{2j} is the reset time of the jth accent command. The F_0 model also makes use of other parameters (time constants α_i and β_j) to express functions G_{pi} and G_{aj}, but, in the current experiments, they are respectively fixed at 3.0 s^{-1} and 15.0 s^{-1} based on the former F_0 contour analysis results.

6. Estimation of F_0 model parameters

6.1. Input and output parameters

In our original methods for F_0 model parameter estimation, taking the fact that the estimation of F_0 model parameters is done in accent phrase basis, input parameters were selected from those related only to the accent phrase in question. However, they do not include direct information on how the accent phrase is related with other accent phrases in a sentence. In the new methods, we added a code to indicate the depth of "bunsetsu" boundary between current and preceding accent phrases, which was obtainable by the Japanese text parser KNP for syntactic analysis [8]. Figure 1 indicates the analysis result by the parser for the sentence "arayuru geNjitsuo subete jibuNno hoHe nejimagetanoda ([He] twisted all the reality to his side.)." In the example, "bunsetsu's" and accent phrases are the same, and the boundary depth codes are obtained by simply shifting the distances rightward. We further changed par-of-speech categories to those obtained form the Japanese morpheme analysis system JUMAN [9], so that we could use JUMAN as the morpheme analyzer of TTS systems. The input parameters of the new methods are summarized in Table 3 with their category numbers. The output parameters for each accent phrase are a set of F_0 model parameters (magnitudes/amplitudes and timings) and a binary flag indicating the existence/absence of a phrase command at the head of the accent phrase. There is no change form the original methods to the new methods. The output parameters are also listed in Table 3. In the table, T_{0off} is the offset of T_0 with respect to the segmental beginning of the accent phrase. T_{1off} and T_{2off} are respectively offsets of T_1 and T_2 with respect to segmental anchor points, which are respectively defined as the beginning of the first high mora (basic unit of Japanese pronunciation mostly coincide with a syllable) for T_1, and the end of the mora containing the accent nucleus for T_2. The first high mora of the accent phrase is either the first mora for accent phrases of type 1 accent, or the second mora for accent phrases of other accent types.

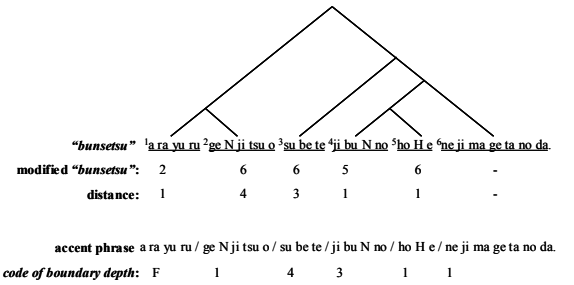

Figure 1: Result of syntactic analysis by KNP and code showing boundary depth for each accent phrase boundary.

Table 3: Input and output parameters for F_0 model parameter estimation for the new methods. "Code of boundary depth" is newly added.

Accent Phrase Features		Category
Input Parameters	Position of Accent Phrase in Sentence	18
	Number of Morae	15
	Accent Type	10
	Number of Words	7
	Part-of-Speech of the First Word	14
	Subsidiary Part-of-Speech of the First Word	11
	Conjugation Type of the First Word	28
	Part-of-Speech of the Last Word	14
	Subsidiary Part-of-Speech of the Last Word	11
	Conjugation Type of the Last Word	28
	Code of Boundary Depth	11
Output Parameters	Flag of Phrase Command (*PF*)	2 (1 or 0)
	Phrase Command Magnitude (A_p)	Continuous
	Offset of T_0 (T_{0off})	Continuous
	Accent Command Amplitude (A_a)	Continuous
	Offset of T_1 (T_{1off})	Continuous
	Offset of T_2 (T_{2off})	Continuous

6.2. Experiments

The prosodic corpus used for the experiments contains 503 sentence utterances by the male speaker MHT included in ATR's continuous speech corpus [10]. It was divided into three parts in the same way as indicated in section 4: 388 sentences (2803 accent phrases) used as training data, 48 sentences (262 accent phrases) used as test data, and 50 sentences used as validation data for neural networks. The F_0 model parameters for the training data were derived from J-ToBI labels attached to the corpus already. First, timing parameters were estimated using J-ToBI labels as suggested in [11], and, then, the analysis-by-synthesis process was carried out for F_0 contours extracted form the speech waveform. The value of F_{0min} was fixed to 51.0 Hz.

The division into accent phrases, as well as the information related to accent types, was also derived from J-ToBI labels. Mora boundaries were obtained from the original phoneme

boundaries using simple rules. In the current experiments, division into accent phrases, as well as the information related to accent types were also derived from the J-ToBI labels.

The results are summarized in Table 4, where mean square errors (MSE's) on logarithmic scale between synthesized F_0 contours using the estimated model parameters and observed F_0 contours are listed as indices of estimation performance. When synthesizing F_0 contours, information on segmental timing and voiced/unvoiced of the target speech is utilized as it is. The MSE values in the table are those averaged over all the test sentences. The table also contains the MSE when command values estimated form the J-ToBI labels are used (target F_0 contours). Clearly better results were obtained by the new methods. Table 5 shows the multiple correlation coefficients for MLRA. Although the improvements are observable for all the output parameters from the original to the new methods, they are significant for those of phrase components. This result is quite natural, since the syntactic structure of a sentence mostly related to the phrasing.

Table 4: Average mean square errors (MSE's) for 48 test sentences between F_0 contours generated using estimated F_0 parameter values and observed F_0 contours. MSE is calculated as mean square distance (per a frame) between synthesized F0 contour and that of natural utterance.

Methods		MSE
Target		0.127
NN (Original Methods)	MLP-10	0.218
	MLP-20	0.217
	Jordan-10	0.220
	Jordan-20	0.215
	Elman-10	0.214
	Elman-20	0.232
BDT-30	Original	0.226
	New	0.197
BDT-50	Original	0.228
	New	0.201
BDT-70	Original	0.226
	New	0.198
MLRA	Original	0.223
	New	0.198

Table 5: Multiple correlation coefficients of the training data for multiple linear regression analysis.

Output Parameters	Original	New
PF	0.602	0.685
A_p	0.642	0.729
T_{0off}	0.590	0.641
A_a	0.441	0.465
T_{1off}	0.440	0.442
T_{2off}	0.428	0.438

7. Conclusion

Data-driven F_0 contour synthesis scheme under the F_0 model constraints was developed. As for the prediction modules, neural networks, and modules based on binary decision tree and multiple linear regression analysis were tested. A code to indicate the depth of boundary between current and preceding accent phrases was added to the input parameters of predictors. Through experiment it was shown the addition improved the estimation performance especially for phrase components. We also applied the statistical methods to the accent phrase boundary estimation from the input text successfully.

One of the problems of the current scheme is that the prosodic features of the neighboring phrases are not taken into account for the estimation of F_0 model parameters of the current accent phrase. We are now planning to incorporate information of preceding and following accent phrases in the process. Also evaluation experiments for the speech obtained by TTS conversion process is planned.

8. References

[1] Tokuda, K.; Masuko, T.; Miyazaki, N.; Kobayashi, T., 1999. Hidden Markov models based on multispace probability distribution for pitch pattern modeling. *ICASSP*, 229-232.

[2] Fujisaki, H.; Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoust. Soc. Japan*, 5(4), 233-242.

[3] Hirose, K.; Fujisaki, H., 1993. A system for the synthesis of high-quality speech from texts on general weather conditions. *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, E76-A(11), 1971-1980.

[4] Hirose, K.; Eto, M.; Minematsu, N.; Sakurai, A., 2001. Corpus-based synthesis of fundamental frequency contours based on a generation process model. *EUROSPEECH, Aalborg,* 2255-2258.

[5] Hirai, T.; Iwahashi, N.; Higuchi, N.; Sagisaka, Y., 1996. Automatic extraction of F_0 control rules using statistical analysis. *in Advances in Speech Synthesis, Springer,* 333-346.

[6] University of Stuttgart, 1995. Stuttgart neural network simulator -User manual-Version 4.1-, Report No. 6/95.

[7] Edinburgh University, Edinburgh Speech Tools Library – Wagon, http://www.cstr.ed.ac.uk/projects/speeech_tools/manual.

[8] Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.

[9] Kyoto University, Japanese Morpheme Analysis System JUMAN http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.

[10] Speech Corpus Set B. http://www.red.atr.co.jp/database_page/digdb.html

[11] Hirai, T.; Higuchi, N., 1998. Automatic extraction of the Fujisaki model parameters using the labels of Japanese tone and break indices (J-ToBI) system, *Trans. Institute of Electronics, Information and Communication Engineers,* J81-D-II, 1058-1064.