AN F0 CONTOUR CONTROL MODEL FOR TOTALLY SPEAKER DRIVEN TEXT TO SPEECH SYSTEM

Takehiko KAGOSHIMA, Masahiro MORITA, Shigenobu SETO and Masami AKAMINE
Toshiba corporation, kansai research laboratories

ABSTRACT

Totally Speaker Driven Text to Speech System produces high quality and natural speech resembling the acoustic and prosodic characteristics of the original speech corpus. In the F0 contour control of this system, an F0 contour of a whole sentence is produced by concatenating segmental F0 contours generated by modifying vectors that are representatives of typical F0 contours. The representative vectors are selected from the F0 contour codebook, which is designed so as to minimize the approximation error between F0 contours generated by the proposed model and real F0 contours extracted from a speech corpus. It was confirmed by experiments with Japanese speech corpus that F0 contours can be modeled with small approximation errors by only 48 representative vectors, and the synthetic speech sounded very natural and resembled the prosodic characteristics of the original speaker.

1. INTRODUCTION

Several automatic learning techniques that derive parameters of F0 contour control models from a speech corpus have recently been proposed to improve synthetic speech quality and naturalness [1][2]. While these corpus-driven approaches have potential to produce fairly good speech quality, there is still a gap between what people expect from TTS and what these conventional methods can deliver.

The authors have developed Totally Speaker Driven Text to Speech System[3][4], where parameters of an F0 contour control model and speech synthesis units for a synthesizer are automatically derived from a speech corpus in order to synthesize high quality and natural speech that resembles the acoustic and prosodic characteristics of the original speech corpus. An F0 contour control model of the system is simple to realize because only two parameters, a representative vector and an offset level, are required for each phrase. Local F0 contours are expressed by patterns of the representative vectors which compose F0 contour codebook, and a rough F0 contour of the whole sentence is expressed by the offset level of each phrase. The F0 contour codebook is designed so as to minimize the approximation error by a training method similar to closed-loop training on vector quantization. Thus, in spite of the simplicity, this model can precisely approximate F0 contours of real speech with only scores of representative vectors. Moreover, vector selection rule and offset level prediction rule are also automatically derived by statistical analysis using the result of the codebook training.

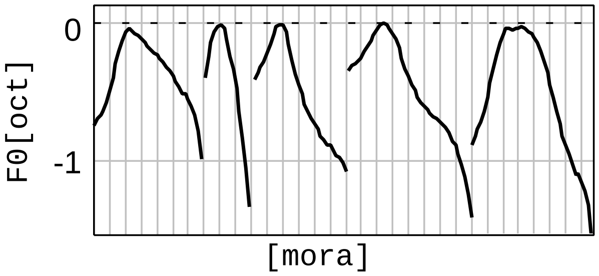


Figure 1: Examples of representative vectors.

2. F0 CONTOUR CONTROL MODEL

An F0 contour of a whole sentence is produced by concatenating F0 contours of phrases, which are called segmental contours. The segmental contours are generated by modifying vectors that are representative of typical F0 contours in the original speech. Figure 1 shows examples of representative vectors, which were produced from a Japanese speech corpus. The vertical and horizontal axes indicate logarithmic frequency and mora respectively. The vectors are normalized as the dimensions per mora is constant. If the number of morae of a phrase is smaller than that of an representative vector, only the front part of the vector is used.

The process of generating segmental contours consists of the following three phases:

1. A representative vector for each phrase is selected from an F0 contour codebook.
2. Each mora of the representative vector is expanded or contracted according to given duration.
3. The representative vector is translated on the logarithmic frequency axis.

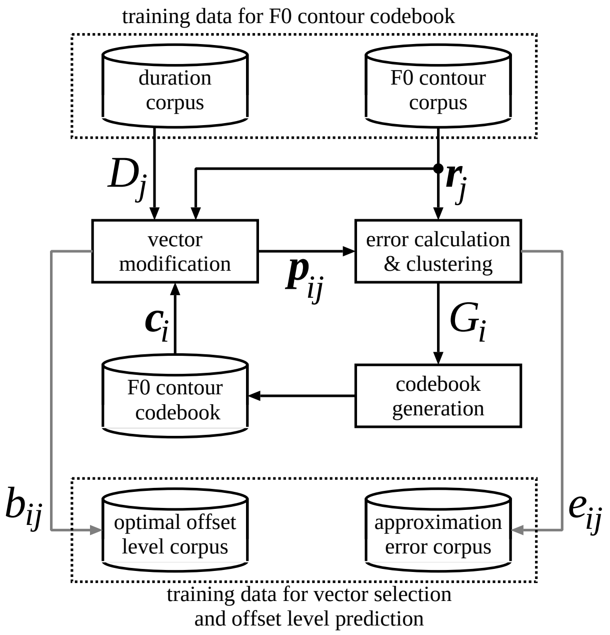

training data for F0 contour codebook
duration corpus
F0 contour corpus
D_j
r_j
vector modification
p_{ij}
error calculation & clustering
c_i
G_i
F0 contour codebook
codebook generation
b_{ij}
e_{ij}
optimal offset level corpus
approximation error corpus
training data for vector selection and offset level prediction

Figure 2: Training of an F0 contour codebook.

The representative vector selection and the prediction of the offset level are performed based on information given by text analysis.

A representative vector $\boldsymbol{c}$, an offset level b and a matrix D, which acts as expansion and/or contraction, can describe a segmental contour $\boldsymbol{p}$ as follows:

$$\boldsymbol{p} = D\boldsymbol{c} + b\boldsymbol{i}, \tag{1}$$

where $\boldsymbol{i}$ represents the vector in which all the elements are one.

3. TRAINING OF AN F0 CONTOUR CODEBOOK

An F0 contour codebook is trained by using a segmental contour corpus and a duration corpus, which are extracted from large speech corpus of a single speaker. Figure 2 shows the process of the codebook training. First, a segmental contour $\boldsymbol{p}_{ij}$ is generated by modifying a representative vector $\boldsymbol{c}_i$ so as to approximate training data $\boldsymbol{r}_j$. This segmental contour generation is performed for combinations of all the representative vectors $\boldsymbol{c}_i$ $(i = 1, 2, \cdots N_c)$ and all the training data $\boldsymbol{r}_j$ $(i = 1, 2, \cdots N_r)$. Secondly, approximation errors e_{ij} are calculated between $\boldsymbol{r}_j$ and $\boldsymbol{p}_{ij}$. According to the approximation error, the training data $\boldsymbol{r}_j$ are grouped to clusters G_i where each cluster is represented by $\boldsymbol{c}_i$. Lastly, every representative vector is renewed so that the total approximation error in each cluster is minimized. This process is repeated until the sum of the total approximation errors in all clusters converges.

3.1. Clustering

The approximation error e_{ij} between $\boldsymbol{r}_j$ and $\boldsymbol{p}_{ij}$ is defined as squared error:

$$\begin{aligned} e_{ij} &= (\boldsymbol{r}_j - \boldsymbol{p}_{ij})^T(\boldsymbol{r}_j - \boldsymbol{p}_{ij}) & (2)\\ &= (\boldsymbol{r}_j - D_j\boldsymbol{c}_i + b_{ij}\boldsymbol{i})^T(\boldsymbol{r}_j - D_j\boldsymbol{c}_i + b_{ij}\boldsymbol{i}), & (3)\end{aligned}$$

where b_{ij} is the optimal offset level that minimizes the approximation error, and it is obtained as follows:

$$b_{ij} = \frac{\boldsymbol{i}^T(\boldsymbol{r}_j - D_j\boldsymbol{c}_i)}{\boldsymbol{i}^T\boldsymbol{i}}. \tag{4}$$

The training data $\boldsymbol{r}_j$ are grouped to cluster G_i as follows:

$$G_i = \{\boldsymbol{r}_j | \min[e_{1j}, \cdots, e_{nj}] = e_{ij}\}, \tag{5}$$

where $\min[x_1, \cdots, x_n]$ indicates the minimum value among $x_1, \cdots, x_n$.

When the sum of the total approximation errors converges and the codebook training is finished, the latest optimal offset levels $b_j = b_{ij}$ $(j = 1, 2, \cdots N_r, \boldsymbol{r}_j \in G_i)$ and the approximation errors e_{ij} $(i = 1, 2, \cdots, N_c, j = 1, 2, \cdots, N_r)$ are stored as training data for representative vector selection and offset level prediction described below.

3.2. Codebook generation

The total approximation error in the cluster G_i is represented by E_i:

$$E_i = \sum_{\boldsymbol{r}_j \in G_i} (\boldsymbol{r}_j - D_j\boldsymbol{c}_i + b_{ij}\boldsymbol{i})^T(\boldsymbol{r}_j - D_j\boldsymbol{c}_i + b_{ij}\boldsymbol{i}) \tag{6}$$

The representative vector $\boldsymbol{c}_i$ that minimizes E_i can be obtained by solving the following equation:

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{c}_i} &= 0 & (7)\\ (\sum D_j^T D_j)\boldsymbol{c}_i &= \sum D_j^T(\boldsymbol{r}_j - b_{ij}\boldsymbol{i}). & (8)\end{aligned}$$

4. REPRESENTATIVE VECTOR SELECTION

From the codebook, a representative vector is selected for each phrase to form the F0 contour for the whole sentence. To select an optimal representative vector for each phrase, each vector is evaluated in terms of predicted approximation error, and a vector with the minimal error is selected. The approximation error is predicted according to grammatical attributes obtained by text analysis using Quantification Method Type I. The predicted error $\hat{e}_{ij}$ is formulated as follows:

$$\hat{e}_{ij} = \sum_{k}\sum_{m} a_{ikm}\delta_j(k,m), \tag{9}$$

where $\delta_j()$ is the characteristic function:

$$\delta_j(k,m) = \begin{cases} 1 & \cdots \text{ if j-th sample falls into} \\ & \quad\ \text{category } m \text{ of attribute } k \\ 0 & \cdots \text{ otherwise} \end{cases}. \tag{10}$$

The attributes which compose the function $\delta_j(k,m)$ represent linguistic characteristics such as phrase length, accent type, the part of speech and so on. The coefficients a_{ikm} of the prediction model are determined so as to minimize

$$\sum_{j=1}^{N_r}(e_{ij} - \hat{e}_{ij})^2, \tag{11}$$

where e_{ij} are real approximation errors that have been obtained through training of the codebook.

5. OFFSET LEVEL PREDICTION

An offset level is predicted for each phrase by the same prediction model as the approximation error prediction above. The coefficients of the prediction model are obtained by minimizing

$$\sum_{j=1}^{N_r}(b_j - \hat{b}_j)^2, \tag{12}$$

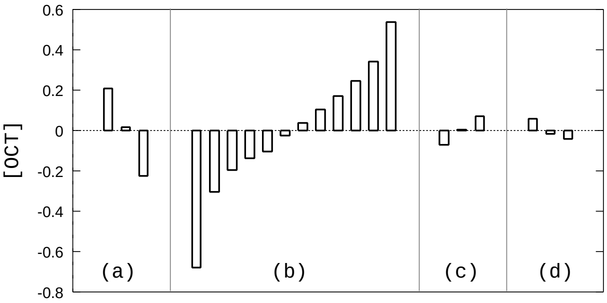
where $\hat{b}_j$ are predicted offset levels, and b_j are the optimal levels that were employed for modification of a representative vector in training of the codebook.

6. EXPERIMENT

6.1. Data sets

A Japanese speech corpus uttered by one male professional narrator was used for preparing a segmental contour corpus and a duration corpus. The speech corpus includes 866 sentences (6398 phrases). F0 values were automatically generated, and unvoiced portions were filled by interpolation. The interval between analysis point is 10.0 msec. Phoneme labels were produced by HMM-based labeling system. Linguistic characteristics for prediction of approximation errors and offset levels were extracted by text analysis and incorrect data were manually corrected. The grammatical attributes were composed of the undermentioned factors of current, preceding and following phrases:

1. position of the phrase in the sentence.



(a) position of current phrase in a sentence (first/middle/last)
(b) current phrase length (1/2/3/4/5/6/7/8/9/10/11/12)
(c) the phrase which current phrase connect to (next/next but one/next but over one)
(d) the phrase which preceding phrase connect to (next/next but one/next but over one)

Figure 3: Coefficients of an offset level prediction model.

2. phrase length.
3. accent type.
4. stress.
5. the phrase which the current phrase connects to.
6. the part of speech.
7. inflection.
8. the connection between the preceding phrase and the current phrase.

6.2. Experimental results

An F0 contour codebook, a vector selection rule and an offset level prediction rule were produced by the proposed method. The F0 contour codebook was composed of 48 representative vectors. Table 1 shows the RMS errors between original F0 contours in training data and those generated by the proposed F0 contour control model with the optimal offset levels. Figure 3 shows coefficients of offset level prediction model according to factors with large contribution. RMS error between the optimal offset levels and predicted offset levels are given by cross validation as shown in table 2.

F0 contours of ten sentences, which were not included in training data, were produced by the proposed method, provided that results of text analysis were correct. RMS error of these ten F0 contours was 0.270[oct]. Figure 4 shows one of the produced F0 contours.

To confirm the validity of the proposed method, listening tests were carried out. Synthetic speeches of seven sentences with proposed method and those with conventional method[5] were produced by the same

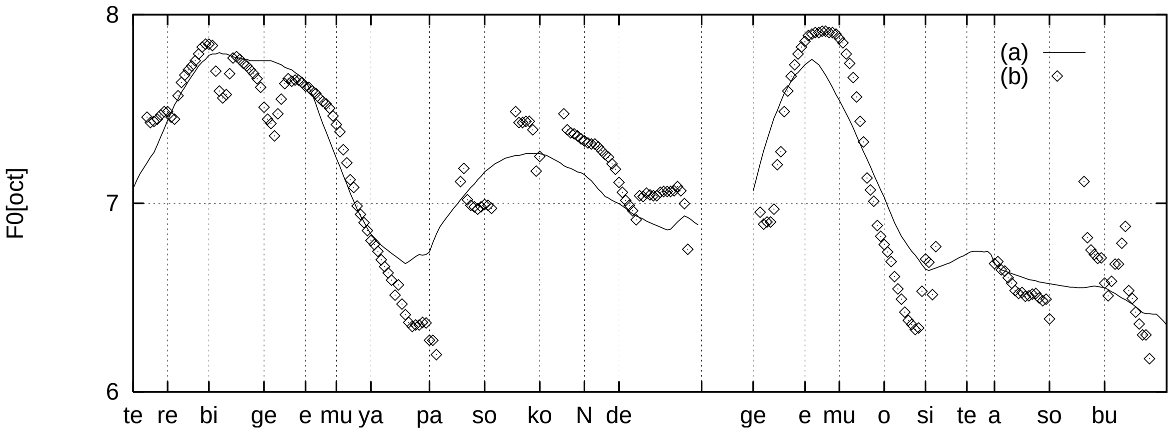


Figure 4: An example of an F0 contour((a)proposed method (b)original).

Table 1: Approximation errors.

with the optimal vectors	0.126 [oct]
with vectors selected by rule	0.167 [oct]

Table 2: Prediction errors of offset levels.

training data	0.177 [oct]
test data	0.185 [oct]

Table 3: Results of listening tests.

	preference score
naturalness test	84.3 %
individualities test	92.9 %

synthesizer. 5 subjects compared them on two listening tests. One was a preference test to evaluate naturalness, and the other was a test where subjects selected a synthetic speech that resembles the speech synthesized with the original prosody from a viewpoint of speaker individualities. Table 3 shows the results of the tests.

7. CONCLUSION

The F0 contour codebook, representative vector selection rule and offset level prediction rule can be designed so as to minimize approximation errors between F0 contour of original speech corpus and produced F0 contours by the proposed method. The vector selection method is based on approximation error of each vector and not selection rate of the optimal vector. Therefore, the proposed method is so robust that vectors with critical errors are hardly selected. It was confirmed by experiments with Japanese speech corpus that F0 contours can be modeled with small approximation errors by only 48 representative vectors, and vector selection and offset level prediction were relatively precise. Thus the synthetic speech sounded very natural and resembled the prosodic characteristics of the original speaker. It is possible to apply the proposed method to other languages and other speech styles because the F0 contour codebook, the vector selection rule and the offset level prediction rule are automatically derived from speech corpus independent of linguistic rules. Therefore, the proposed method makes it easy to develop TTS system for a new speaker.

8. REFERENCES

1. Hirai, T., Iwahashi, N., Higuchi, N., and Sagisaka, Y. “Automatic Extraction of Fundamental Frequency Control Rules Using Statistical Analysis,” (in Japanese) Trans. IEICE vol. J78-D-II, no. 11, pp.1572-1580, Nov.,1995.
2. Huang, X., Acero, A., Hon, H., Ju, Y., Liu, J., Meredith, S., and Plumpe, M. “Recent improvements on Microsoft's trainable text-to-speech system - Whistler,” Proc. IEEE ICASSP'96, pp.959-962, April 1996.
3. Akamine, M., and Kagoshima, T. “Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Speech System (TOS Drive TTS),” Proc. ICSLP'98, Nov. 1998.
4. Seto, S., Morita, M., Kagoshima, T., and Akamine, M. “Automatic Rule Generation for Linguistic Features Analysis Using Inductive Learning Technique,” Proc. ICSLP'98, Nov. 1998.
5. Shiga, Y., Hara, Y., and Nitta, T. “A Novel Segment-Concatenation Algorithm for a Cepstrum-Based Synthesizer,” Proc. ICSLP'94, pp.1783-1786, Sep. 1994.