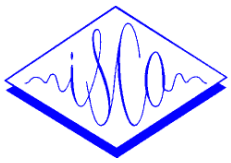

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# A STUDY ON PITCH PATTERN GENERATION USING HMM-BASED STATISTICAL INFORMATION

*Toshiaki FUKADA, Yasuhiro KOMORI, Takashi ASO and Yasunori OHORA*

Media Technology Laboratory, Canon Inc., Kawasaki, 211 JAPAN (toshiaki@cis.canon.co.jp)

## ABSTRACT

This paper describes a novel pitch pattern generation method for speech synthesis using Hidden Markov Models (HMMs). In the proposed method, the $F_0$ contours of minor phrase are modeled by HMMs (pitch-HMMs). The pitch-HMMs are trained using $F_0$ and $\Delta F_0$ considering phonetic environments (e.g. accent type, mora count, mora position, phonemic category, etc.). To evaluate the pitch-HMMs, accent identification experiments are performed. The results indicate that the pitch-HMMs can capture the movement in $F_0$ contours appropriately. In the $F_0$ contour generation experiments, the proposed method yields an averaged root mean square error of 132cent (equivalent to 9.2Hz at 120Hz) between the original and the generated $F_0$ contours. Furthermore, an application of the proposed method to text-to-speech system is also discussed.

## 1. INTRODUCTION

A good generation model of fundamental frequency ($F_0$ contours) is essential for speech systems. Recently, several methods of $F_0$ contour modeling based on statistical analysis have been proposed [1] [2] [3] [4] [5]. Especially Hidden Markov Model (HMM) is well-known as a powerful method for speech systems. However, most of $F_0$ contour modelings based on HMM have been proposed on the study of speech recognition for word accent identification[3] and phrase boundary detection [4], while few study of speech synthesis was investigated [5]. In [5], additionally, its study was insufficient for applying the prosodic HMM to text-to-speech systems [1].

In this paper, we propose a new $F_0$ contour generation using phone-based pitch-HMM[6] aiming at text-to-speech. The proposed method has the following advantages.

- By taking account of phonetic environments (e.g. accent type, mora count, mora position, phonemic category, etc.), precise models can be obtained.
- Detail movement in $F_0$ contours can be captured by several numbers of states in the HMM.
- Not only $F_0$ but its differential ($\Delta F_0$) can be used directly and simultaneously.

In the following sections, we first describe $F_0$ contour modeling using HMMs. Section 3 then presents a method of $F_0$ contour generation using pitch-HMMs. Experiments of word accent identification and $F_0$ contour generation for isolated words are both described in section 4. Section 5 discusses how to apply the proposed method to text-to-speech.

[1] Only monosyllabic words which did not include nasals and glides were tested.

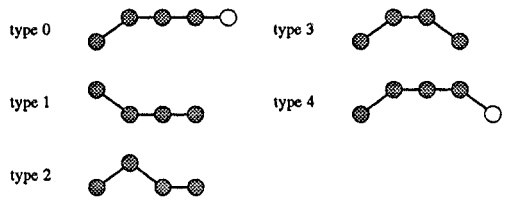


Figure 1: Subjective pitch patterns of 4-mora word.

## 2. $F_0$ CONTOUR GENERATION USING PITCH-HMMS

### 2.1 Accent Types in Japanese

Spoken Japanese consists of a chain of minor phrases ("prosodic words" defined by Fujisaki[7]). An $n$-mora minor phrase of Japanese in the Tokyo dialect can be classified into $(n+1)$ accent types, which are usually denoted by "type $i$" accents ($i = 0$ to $n$). Figure 1 shows examples of subjective pitch patterns of 4-mora word. In this paper, type 0 and type $n$ are treated as the same category because they are indistinguishable in isolated word utterances.

### 2.2 Hidden Markov Modeling

$F_0$ contours of minor phrase are modeled by concatenating pitch-HMMs. They are trained by considering phonetic environments (e.g. accent type, mora count, mora position, phonemic category, etc.). Therefore pitch-HMMs can express the difference of $F_0$ contours related to phonemes[8]. It is also reported that expressing the microscopic movement in a phoneme improves the intelligibility of synthesized speech[9]. This movement in $F_0$ contours can be captured by the states in the pitch-HMM. In addition, plural prosodic parameters can be used for training. This helps to make the model more stable.

### 2.3 Prosodic Parameters for Pitch-HMMs

The following two features are used for pitch-HMM training.

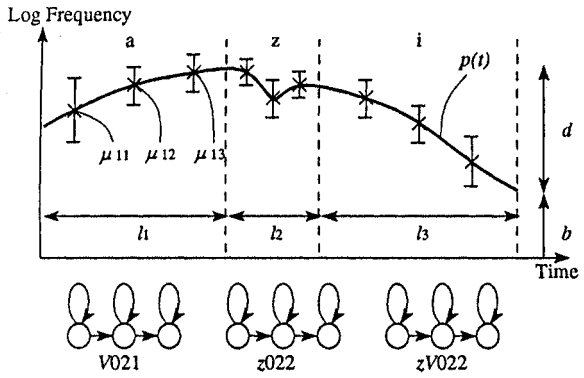


Figure 2: Conceptual figure of $F_0$ contour generation. The *V021* means that *V(owel)* for phonemic category, 0 for accent type, 2 for mora count and 1 for current mora position.

- $F_0$ : Since raw $F_0$ value, $F_{0raw}$ may vary widely among words, we normalize $F_0$ contours within each word. The normalized $F_0$ is given by the following equations.

$$F_0 = (F_{0raw} - b)/d \quad (1)$$

where $b$ is a bias which is the minimum $F_0$ value in the $F_0$ contours and $d$ is a dynamic range which is the difference between the maximum and the minimum in the $F_0$ contours (see Figure2).

- $\Delta F_0$ : Differenced fundamental frequency (henceforth $\Delta F_0$) provides information about relative changes in $F_0$ contours.

### 2.4 $F_0$ Contour Generation

$F_0$ contours are generated by concatenating the desired pitch-HMMs in the following steps. (See Figure 2; the example is "azi(/*taste*/)".)

1. select pitch-HMMs according to the phonetic environments (*V*021, *z*022, *zV*022)
2. align the pitch-HMMs according to the segmental durations ($l_1$, $l_2$, $l_3$)
3. determine target pitch points using the mean values of the pitch-HMMs ($\mu_{11}$, $\mu_{12}$, $\mu_{13}$, etc.)
4. interpolate the target pitch points ($p'(t)$)
5. multiply a dynamic range ($d$) and add a bias ($b$)

$$p(t) = d \cdot p'(t) + b \quad (2)$$

Considering application to text-to-speech system, the segmental durations and the values of bias and dynamic range have to be determined (see section 4).

## 3. EXPERIMENTS

### 3.1 Training of Pitch-HMMs

**A. Speech Material** For pitch-HMM training and for carrying out experiments, we used 5,240 words sampled at 12kHz in the ATR database[10] uttered by one male speaker. The database contains phoneme boundary information manually segmented. Lists of the accent types and the mora counts for these words are also prepared.

Table I: Types of phonemic category.

| model | class of pitch-HMMs |
|---|---|
| rough (2 classes) | vowels : *V<br>voiced consonants : * |
| detailed (15 classes) | vowels : V(no consonant), ptkV, fshV, bdgV, mnV, rV, wV, yV, zV<br>voiced consonants : bdg, mn, r, w, y, z |

The "ptkV" indicates vowels preceding consonant class "ptk".

**B. Method of Analysis** The $F_0$ contour estimation procedure was based upon the FFT cepstral analysis. We computed the fundamental frequency every 2.5 ms, or 30 speech samples. The fundamental frequency was warped on a logarithmic scale, and then smoothed out. In this study, hand correction of extraction errors was not performed. The first regression coefficients of the $F_0$ were estimated every 77.5ms for $\Delta F_0$.

**C. Training** Accent type, mora count, mora position and phonemic category were considered as phonetic environments. Pitch-HMMs are single Gaussian density HMM with 3 states. They were trained by Forward-Backward algorithm.

It is widely known that $F_0$ contours differ according to phonemic environments, even if the words have the same accent type and the same mora count[8]. Considering these phenomena, it can be expected that the precision of $F_0$ contour estimation may be improved by increasing the number of pitch-HMMs taking account of phonemic environments.

From this point of view, we selected two HMM sets, which consist of 2 classes(*rough models*) and 15 classes(*detailed models*), shown in Table I.

**D. Other Conditions** The pitch-HMMs can be aligned appropriately using their transition probabilities according to the segmental durations. However in the following experiments, the pitch-HMMs were simply aligned at the positions of 1/6, 3/6, 5/6 in each phoneme. As for the interpolation, although various techniques (e.g. spline interpolation or Bézier curve) can be applied, the simplest straight line interpolation was adopted in this study.

### 3.2 Accent Identification Experiments

For investigating how the pitch-HMMs can express the accent type appropriately, we evaluated by accent identification experiment using 4 mora words (1,565 words) among the 5,240 words[2]. This experiment was simply performed using only vowel parts divided by null code. The accent pattern models were created by concatenating the null code between the rough pitch-HMMs (see Figure 3). Viterbi beam search was adopted for the evaluation.

The experimental results are shown in Table II and Table III indicating the effectiveness of $\Delta F_0$ for the pitch-HMMs. Drastic improvement from 82.4% to 91.5% ac-

[2]Speech data which had voiceless sound of vowel and concatenated phoneme labels were taken off.

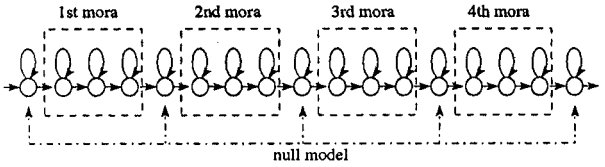


Figure 3: Accent pattern modeling for identification.

Table II: Confusion Matrix of Accent Identification of 4 mora words. Pitch-HMMs are trained by using two features, $F_0$ and $\Delta F_0$.

| Accent type | 0 | 1 | 2 | 3 | ID rate |
|---|---|---|---|---|---|
| 0 | 921 | 4 | 50 | 76 | 87.6 % |
| 1 | 0 | 143 | 4 | 0 | 97.3 % |
| 2 | 3 | 0 | 78 | 6 | 89.7 % |
| 3 | 4 | 0 | 20 | 256 | 91.4 % |

(Total ID rate : 91.5 %)

Table III: Confusion Matrix of Accent Identification of 4 mora words. Pitch-HMMs are trained by using one feature, $F_0$.

| Accent type | 0 | 1 | 2 | 3 | ID rate |
|---|---|---|---|---|---|
| 0 | 862 | 6 | 44 | 139 | 82.0 % |
| 1 | 2 | 144 | 1 | 0 | 98.0 % |
| 2 | 11 | 0 | 65 | 11 | 74.7 % |
| 3 | 48 | 0 | 22 | 210 | 75.0 % |

(Total ID rate : 82.4 %)

curacy was obtained. Furthermore, the result with $\Delta F_0$ indicates that the pitch-HMMs express appropriate accent patterns.

### 3.3 RMSE Experiments

We also measured an average of root mean square error (RMSE) between the original $F_0$ contours and the one generated by the proposed method using the 5,240 words. The segmental durations and the values of the dynamic range and the bias were given by the original.

**A. Rough Model** As a result, the RMSE using rough models was 146cent (equivalent to 10.2Hz at 120Hz). Figure 4 shows the examples of $F_0$ contour generation.

**B. Detailed Model** The RMSE came up to 132cent (equivalent to 9.2Hz at 120Hz). The detailed models yielded improvement of about 10 % as compared with the rough models. Figure 5 shows the examples of $F_0$ contour generation. As shown in Figure 5, $F_0$ contours were well estimated in comparison with Figure 4. It is also seen that the microscopic movement in the $F_0$ contours corresponding to phoneme is well captured (see /z/ in Figure 5(b)).

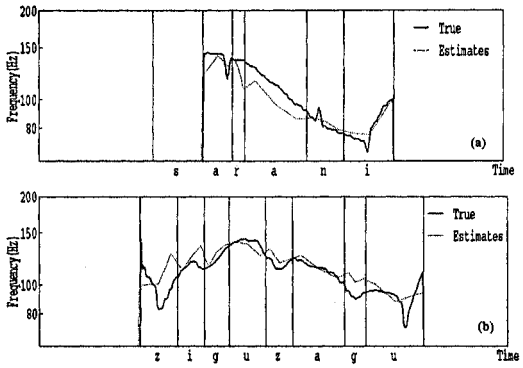


Figure 4: Examples of the $F_0$ contour generation using rough models. (a) "sarani (/moreover/)", (b) "ziguzagu (/zigzag/)".

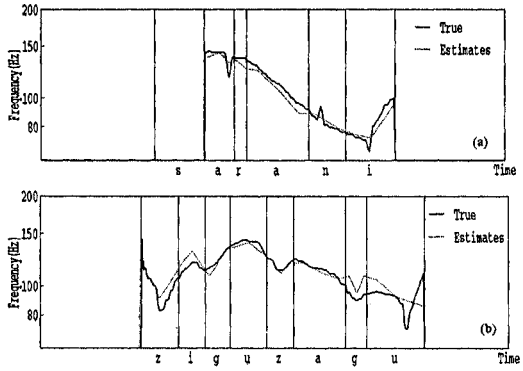


Figure 5: Examples of the $F_0$ contour generation using detailed models.

## 4. APPLICATION TO TEXT-TO-SPEECH

### 4.1 Estimation of Bias and Dynamic Range

Considering applying the proposed method to a text-to-speech system, it is required to estimate values of a bias and a dynamic range for sentences. We applied the categorical multiple regression technique [11] to estimate these values. Values of the bias or the dynamic range are predicted by the following equation.

$$P_i = \sum_{j=1}^{R} \sum_{k=1}^{C_j} a_{jk} \delta_i(jk), \qquad (i = 1, \cdots, N) \qquad (3)$$

where $P_i$ is the predicted value of the $i$-th sample, $R$ is the number of the factors and $C_j$ is the number of the categories of factor $j$. $\delta_i(jk)$ is the characteristics function:

$$\delta_i(jk) = \begin{cases} 1, & \text{if the } i\text{-th samples falls into} \\ & \text{category } k \text{ of factor } j, \\ 0, & \text{otherwise.} \end{cases} \qquad (4)$$

$a_{jk}$ is obtained by minimizing the summation of the prediction error $E$:

$$E = \sum_{i=1}^{n} (p_i - P_i)^2. \qquad (5)$$

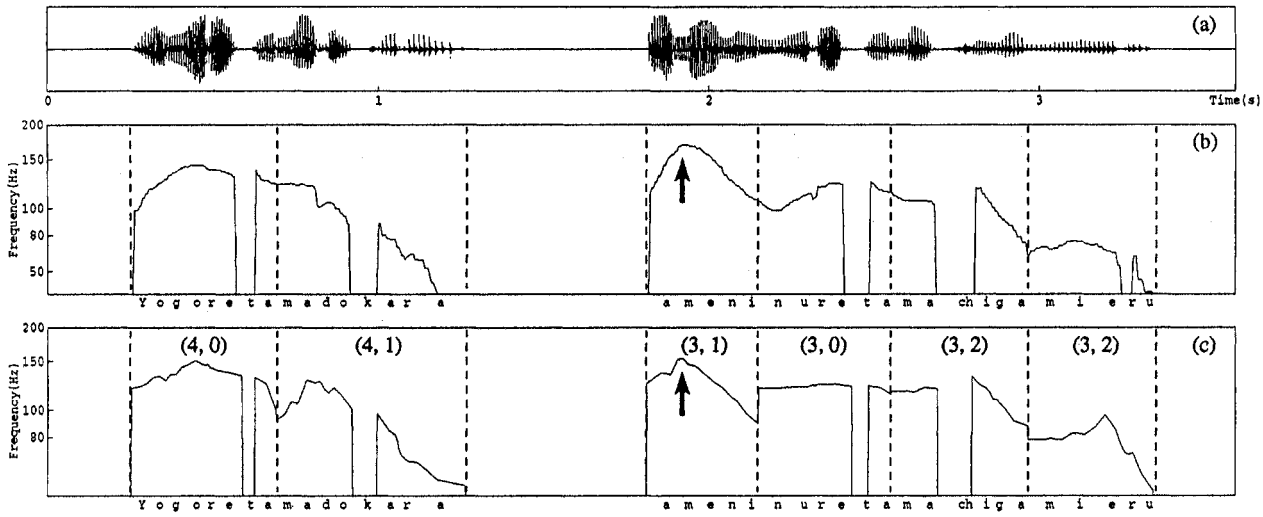


Figure 6: $F_0$ contours for sentence using pitch-HMMs and the categorical multiple regression technique. (a) Waveform, (b) original $F_0$ contours, (c) estimated $F_0$ contours. This sentence consists of 6 minor phrases. (4, 0) indicates the mora count and the accent type.

### 4.2 $F_0$ Contour Generation for Sentence

The bias and the dynamic range were computed from each minor phrase of 503 sentences in the ATR database[10]. The pitch-HMMs were also trained using this database. Here, boundary type, accent type, part of speech, position of top syllable in pause phrase and dependency structure were used as factors for prediction. The maximum values in the minor phrases (bias + dynamic range) were predicted in stead of the values of the biases. This is because we found that the maximum values were more stable.

Figure 6 shows an example of $F_0$ contour estimation for sentence using the pitch-HMMs and the categorical multiple regression technique. In this figure, the original segmental durations were used. As seen in the third minor phrase ("ameni"), the position of the maximum $F_0$ can be well determined, even though the accent type is type 1 (falls). Since the values of the bias or the dynamic range are decided regardless of the relative position, there are several gaps at the minor phrase boundaries. This problem has to be investigated.

## 5. SUMMARY

In this paper, we have presented the new method of HMM-based $F_0$ contour generation. The $F_0$ contours are modeled by the pitch-HMMs. The pitch-HMMs are trained considering the phonetic environments using $F_0$ and $\Delta F_0$. The experiments of accent identification indicated that the pitch-HMMs could capture the movement in $F_0$ contours appropriately. In the $F_0$ contour generation experiments for isolated words, the detailed models classified into 15 phonemic categories achieved the improvement of 10 % for RMSE as compared with the rough models. Furthermore, application to $F_0$ contour generation for text-to-speech by incorporating the categorical multiple regression technique was discussed. From these experimental results, we expect that the proposed method is powerful and useful for text-to-speech system.

## REFERENCES

[1] M. Abe and H. Sato, "Two-stage $F_0$ Control Model Using Syllable Based $F_0$ Units," *Proc. ICASSP-92*, pp.II.53–II.56, 1992.

[2] T. Hirai, N. Iwahashi, N. Higuchi and Y. Sagisaka, "Auto Classification of $F_0$ Control Commands using Statistical Analysis," (*in Japanese*) *IEICE Technical Report*, vol. SP94-12, May. 1994.

[3] T. Yoshimura, S. Hayamizu and K. Tanaka, "Word Accent Patterns Modelling by Concatenation of Mora Hidden Markov Models," *Proc. ICASSP-94*, 11.10, pp.I.69–I.72, 1994.

[4] S. Takahashi and S. Matsunaga, "Stochastic Prosody Modeling for Accent Phrase Boundary Detection in Continuous Speech," (*in Japanese*) *IEICE Technical Report*, vol. SP90-71, Dec. 1990.

[5] A. Ljolje and F. Fallside, "Synthesis of Natural Sounding Pitch Contours in Isolated Utterances Using Hidden Markov Models," *IEEE Trans. Acoust., Speech & Signal Process.*, **ASSP-34**, 5, pp.1074–1080, Oct. 1986.

[6] T. Fukada, Y. Komori, T. Aso and Y. Ohora, "Generation of Word Pitch Pattern Using HMM-based Statistical Information," (*in Japanese*) *Proc. ASJ Spring Meeting*, 2-8-12, pp.229–230, Mar. 1994.

[7] H. Fujisaki, K. Hirose and N. Takahashi, "Manifestation of Linguistic Information in the Voice Fundamental Frequency Contours of Spoken Japanese," *Trans. IEICE*, vol. E76-A, no. 11, pp.1919–1926, Nov. 1993.

[8] H. Sato, "Analysis of Fundamental Frequency Characteristics Related to Phonemes," (*in Japanese*) *Proc. ASJ Fall Meeting*, 2-3-18, pp.259–260, Oct. 1989.

[9] S. Takeda, "A Model for Generating Fundamental Frequency Contours Considering Phonemic Fluctuation and Rules for Speech Synthesis," (*in Japanese*) *Trans. IEICE*, vol. J73-A, no. 3, pp.379–386, Mar. 1990.

[10] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda and H. Kuwabara, "A Large-Scale Japanese Speech Database," *Proc. ISCLP-90*, pp.1089–1092, 1990.

[11] C. Hayashi, "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of View," *Ann. Inst. Math 2*, 1950.