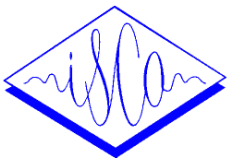

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# MODELING AND GENERATION OF ACCENTUAL PHRASE $F_0$ CONTOURS BASED ON DISCRETE HMMs SYNCHRONIZED AT MORA-UNIT TRANSITIONS

* †*Atsuhiro Sakurai*, ‡†*Koji Iwano*, *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
‡Currently with Tokyo Institute of Technology

## ABSTRACT

We propose a data-driven approach to intonation modeling and generation based on discrete Hidden Markov Models (HMM), where state transitions are synchronized with Japanese rhythmic units called *morae*. Mora-unit $F_0$ contours are encoded using symbols that consist of two codes: the first is an index to a table of stylized mora $F_0$ contours, and the second points to a table of quantized differences of the average $F_0$ contour with respect to the previous mora. Both codebooks contain 32 codes. The HMM is used in generation mode, i.e., it generates a sequence of symbols for an intonational phrase without any input other than the length of the sequence, using a variation of Viterbi search with a modified distance function. In the training phase, the speech database is subdivided into classes according to the attributes of the target accentual phrase, and each class is associated to an HMM. After the output symbol sequence is generated, the $F_0$ contour is constructed using the codebooks and a mora duration pattern. Evaluation experiments show that the HMMs are able to correctly produce $F_0$ contours that reflect their training conditions.

## 1. INTRODUCTION

The use of HMMs in speech synthesis and coding has been proposed in past studies [1][2]. In [1], $F_0$ countours are generated with the help of a random number generator. However, the continuous-valued HMM requires the observations to be artificially reordered after generation. [2] proposes the application of HMMs in very low bit rate speech coding, suggesting again the ability of HMMs to generate outputs. For Japanese, an interesting intonation modeling scheme using HMM is proposed in [3]. HMMs are aligned at the phone level to capture local pitch movements, but the concatenated models cannot deal with phrase-level prosodic structures, and a separate statistical approach based on regression trees is needed to obtain the values of the bias and dynamic range.

More recently, the use of HMM in speech synthesis has been further explored, resulting in a unified solution for segmental and suprasegmental features [4]. In this method, HMMs are used as concatenation units based on cepstral features and delta components. Concatenation of speech units is realized through the concatenation of HMMs, and the best sequence of synthesis-time parameters is found by solving a search problem that resembles the speech recognition paradigm. An advantage of this approach compared to usual waveform concatenation synthesis is that selection of synthesis units is systematic and based on probability scores. In addition, coarticulation phenomena are also dealt with inside the same probabilistic framework. Another advantage is that HMMs can be trained using speech databases and techniques developed for speech recognition.

However, the approach above has limitations when it comes to intonation modeling, especially considering that phone-based models are too short to model long-range suprasegmental phenomena, as occurred in [3]. A separate processing is needed for intonation modeling.

The method proposed in this paper is based on mora-transition discrete HMMs associated to accentual phrases. It intends to take advantage of the perceptual importance shown by Japanese rhythmic units called morae. In addition to the fact that morae are important temporal units, mora-to-mora variation of fundamental frequency is an important perceptual clue to characterize accent types. In addition, it is expected that accentual phrase HMMs are long enough to capture phrase-level prosodic structures.

A similar modeling has been successfully applied to the problem of automatic detection of phrase boundaries in $F_0$ contours [5]. In that work, phrase boundaries are detected as the boundaries between consecutive accentual phrases, represented by the HMMs. Specific models are used to describe different accent types, and different models are assigned to accentual phrases followed by a pause. In this paper, we adapt and apply the models above to $F_0$ contour generation. The models are used in generation mode, i.e., they generate sequences of symbols that reflect the conditions under which they have been trained. The only input to the models is the length of the sequence to be generated, and the best path of state transitions is found using a variation of Viterbi search that generates the most likely sequence of output symbols based on a cost function that takes into account state transition probabilities and output probabilities as well (and also output symbol bigrams in a later

attempt). The generated output symbol sequence is finally converted into an $F_0$ contour for a given number of morae using the codebooks and externally-supplied timing information, and also the value of the $F_0$ baseline.

The following sections describe the method in detail and some evaluation experiments are carried out to verify its validity.

# 2. DESCRIPTION OF THE METHOD

## 2.1. Introduction

This section contains details of the proposed intonation modeling scheme based on discrete HMMs synchronized at mora transitions. First, the database clustering and labeling process is described. Then, we describe the choice of HMM structures, HMM training, and the process of generating output symbols from trained HMMs.

## 2.2. Clustering and Coding

The output of the HMM is formed by 2-code symbols respectively associated to two codebooks: the first code is an index that points to a table of stylized $F_0$ contours, and the second one represents the difference between the average $F_0$ of the current mora with respect to the previous one. In this work, these codes are respectively denominated *shape* code and *delta-$F_0$* code.

The codebooks are obtained using LBG clustering, which is carried out on the $F_0$ contours extracted from 500 sentences of ATR's continuous speech database [6]. Prior to the clustering process, however, mora-unit $F_0$ contours are time-normalized to the duration of 1 second. The $F_0$ contours are also expanded in the vertical direction by the same factor in order to preserve its shape. The *shape* and *delta-$F_0$* codebooks thus obtained contain 32 codes each.

## 2.3. Topology of Accentual-Phrase HMMs

The structure of the discrete HMM is designed based on the number of prosodic events found in the utterance of the accentual phrase for different accent types. For example, accentual phrases of types 0 and 1 are driven basically by two events: a steep rise and decay in type 1, and a slow rise and decay in type 0. Other accent types are basically formed by a rise, a region of slow decay, and a steep decay after the accent nucleus. The observations above make us assume that accentual phrases of type N (N > 1) are roughly characterized by 3 prosodic events, whereas other accent types (0 and 1) are characterized by having 2. Thus, it is reasonable to consider a similar HMM structure for accent types 0 and 1, and a slightly more complex structure for other accent types due to the occurrence of a larger number of prosodic events. In [5], empirical attempts lead to the structures shown in **Figure 1**. Note that initial and final states have been added in the models for convenience.

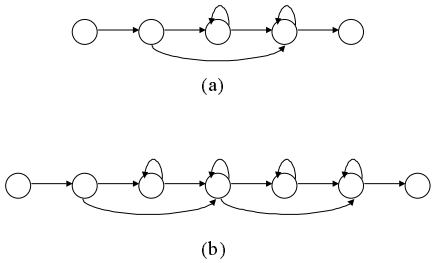


**Figure 1:** HMM for different accent types: (a) types 0 and 1, and (b) other types.

HMMs are used to model the $F_0$ contour of accentual phrases using a method adapted from [5]. In that report, the models are divided by accent types: P0 and P0_P (flat), P1 and P1_P (type 1), PN and PN_P (other accent types), and an additional model P representing pauses. Px and Px_P (x = 0,1,N) differ in that Px_P is followed by a pause. Using techniques similar to speech recognition (Viterbi search and language modeling), these models are recognized within the utterance and prosodic boundaries are found as the boundaries between the accentual phrases represented by the models.

In the present work, however, it is necessary to create models that represent all different variations of accentual phrases needed in speech synthesis. For this reason, models should be created corresponding to several linguistic conditions such as parts-of-speech of component words, phonetic context, number of morae, relative position of the accentual phrase within the sentence, etc., resulting in a much larger number of models. Note, however, that in the experiments that follow we just use two types of linguistic features for model classification: accent type and the position of the accentual phrase within the sentence, in order to cope with limited amount of training data.

## 2.4. HMM Training

After the database is labeled, subdivided according to linguistic attributes of each class, and the classes are associated to their respective HMMs, training is carried out using a conventional forward-backward (FB) algorithm [7], exactly in the same way as in [5]. Two linguistic attributes are used for subdividing the training database, as described in section 3: accent type and relative position of the accentual phrase within the sentence.

## 2.5. Generation of Output Symbol Sequences Based on a Modified Viterbi Algorithm

In speech recognition, the Viterbi algorithm provides a measure of probability (likelihood) of a model having generated a given a sequence of output feature vectors (or discrete symbols in the case of discrete HMMs). In addition, the Viterbi algorithm also produces the state transition path associated to the output sequence with the highest probability, stored by the backtracking function.

In the present system, the objective is to obtain the best state transition path and output symbol sequence given the model and the path length. Upon synthesis, the path length is provided by the linguistic module and corresponds to the number of morae in the accentual phrase. Prior to the Viterbi search, other types of linguistic information would be used in a real-world TTS system to partition the text into accentual phrases, determine their accent types, find their part-of-speech classes, etc., and finally select the appropriate HMM for synthesis.

A Viterbi algorithm is used to find the best state transition path and output symbol sequence. However, there is a difference with respect to the Viterbi search commonly found in speech recognition algorithms. In a typical recognition problem, the sequence of symbols is given and referred to as sequence of observation symbols, whereas in this case it is the very objective of the algorithm to find the sequence of symbols in an optimized way.

In view of that, we modify the distance function of the Viterbi algorithm. In the usual Viterbi algorithm, the distance function $D_{min}$ up to instant $t$ and state $i_t$ is calculated for every allowed transition using the following formula:

$$D_{min}(t,i_t) = D_{min}(t\text{-}1,i_{t\text{-}1}) - log[a(i_t|i_{t\text{-}1})] - log[b(y(t)|i_t)] \quad (1)$$

, where $D_{min}(t\text{-}1,i_{t\text{-}1})$ is the partial distance up to instant $t\text{-}1$, $-log[a(i_t|i_{t\text{-}1})]$ is the state transition likelihood, and $-log[b(y(t)|i_t)]$ is the likelihood of the given output symbol $y(t)$ for state $i_t$.

We can see that in our problem, the likelihood measure $-log[b(y(t)|i_t)]$ cannot be calculated without the output symbol $y(t)$, which is known in a conventional speech recognition task. Two solutions are proposed for this problem. The first is to replace $b(y(t)|i_t)$ by $b(y_{max}(i_t))$, the probability of the output symbol possessing the highest probability at state $i_t$. The output symbol sequence produced by the modified Viterbi algorithm is the highest-likelihood output sequence, along the best state transition path. In this case, the distance function can be expressed as:

$$D_{min}(t,i_t) = D_{min}(t\text{-}1,i_{t\text{-}1}) - log[a(i_t|i_{t\text{-}1})] - log[b(y_{max}(i_t))] \quad (2)$$

A second solution to this problem is to also consider bigram probabilities of output symbols $bigr(y(t)|y(t\text{-}1))$ calculated off-line in addition to the previous solution. The resulting cost function is:

$$D_{min}(t,i_t) = D_{min}(t\text{-}1,i_{t\text{-}1}) - log[a(i_t|i_{t\text{-}1})] - log[b(y_{max}(i_t))] - log\{bigr[y(t)|y(t\text{-}1)]\} \quad (3)$$

Informal observation of $F_0$ contours produced using either distance function reveals that the introduction of bigrams sometimes result in better concatenation of mora-unit $F_0$ contours.

# 3. EVALUATION EXPERIMENTS

## 3.1. Modeling Accent Types

In order to evaluate the validity of the method, we first investigate if the HMMs can correctly learn the position of the accent nucleus from the speech database. We classify all accentual phrases present in the database into accent types, associate them to their corresponding HMMs, and train them on their corresponding data. (Due to limited amount of training data, we just classify accent types into accent types 0,1,2,3, and 'others'). Then, we generate the $F_0$ contours corresponding to each model using the method described in the previous sections for a hypothetical 5-mora accentual phrase with uniform mora duration pattern. **Figure 2** shows the $F_0$ contours generated for accent types 0,1,2, and 3. As can be seen in the figure, the system is able to place the accent nucleus on the correct place for all accent types.

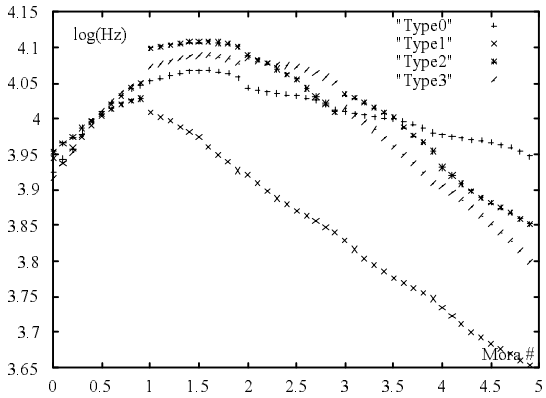


**Figure 2:** Accent type modeling using HMM.

## 3.2. Modeling Prosodic Boundary Levels

In this subsection, we investigate the ability of the HMMs to learn information on prosodic structure. In the experiment that follows, accentual phrases of type 0 in the training database are divided according to the classification of the preceding phrase boundary. Phrase boundaries are classified into 3 levels, according to the J-ToBI break index [8] and the existence or inexistence of pause (see **Table 1**). Three HMMs are then respectively trained on the examples of each prosodic boundary level. After training, $F_0$ contours are generated for the models supposing a 4-mora accentual phrase with a uniform mora duration pattern.

| J-ToBI break index | Existence of pause | Phrase boundary level |
|---|---|---|
| 3 | Yes | 1 |
| 3 | No | 2 |
| 2 | No | 3 |

**Table 1:** Prosodic boundary levels

The generated $F_0$ contours are shown in **Figure 3**. It can be seen that a stronger prosodic boundary produces a more pronounced rise in the $F_0$ contour, as expected from our experience. Note also that the shapes for levels 2 and 3 are slightly deformed, probably due to lack of training data.

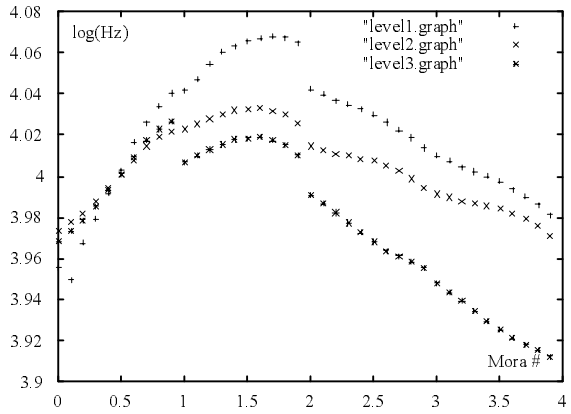


**Figure 3:** Phrase boundary context modeling.

## 4. CONCLUSION

We proposed in this paper a scheme for modeling and generating $F_0$ contours of accentual phrases based on discrete HMMs having mora-synchronized state transitions. The method permits straightforward modeling of arbitrary linguistic constraints, provided that sufficient training data are available. Our evaluation experiments show that the modeling is able to cope with such features as accent type and phrase boundary level. A real implementation in a TTS system would require further training involving several other linguistic classifications, such as phonetic context, syntactic structure, morphological class, etc. The small size of the speech database available for training is a limiting factor to be considered from now on.

## 5. REFERENCES


1. A. Ljolje, F. Fallside, "Synthesis of natural sounding pitch contours in isolated utterances using Hidden Markov Models, " *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 5, pp.1074-1080 (1986-10).

2. E. P. Farges and M. A. Clements, "Hidden Markov Models applied to very low bit rate speech coding, " *IEEE ICASSP'86*, 9.1.1, Tokyo, pp. 433-436 (1986).

3. T. Fukada, Y. Komori, T. Aso, and Y. Ohara, "A study on pitch pattern generation using HMM-based statistical information, " *ICSLP'94*, S14-4.1, pp. 723-726 (1994).

4. K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features, " *IEEE ICASSP'95*, pp. 660-663 (1995).

5. K. Hirose and K. Iwano, "A method of representing fundamental frequency contours of Japanese using statistical models of moraic transition, " *Proc. ESCA Eurospeech'97*, pp. 311-314 (1997).

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

7. Young, S. et. Al., "The HTK Book, version 2.1", Cambridge University (1996).

8. Campbell, N. and Venditti, J., "J-ToBI: An intonation labelling system for Japanese," *Reports of Spring Meeting*, Acoust. Soc. Jpn., pp.317-138 (1995-9).