Stochastic F0 Contour Model Based on the Clustering of F0 Shapes of a Syntactic Unit

Yoichi Yamashita and Tomoyoshi Ishida

Department of Computer Science,
Ritsumeikan University, Japan
{yama,tomoyosi}@slp.cs.ritsumei.ac.jp

Abstract

This paper describes a stochastic modeling between an F0 contour and linguistic features of a sentence for speech synthesis. The F0 contour of a sentence is represented by concatenation of the F0 patterns of a Japanese syntactic unit, *bunsetsu*. A *bunsetsu* F0 pattern is composed of the F0 average and the F0 shape. The F0 average is independently predicted for each *bunsetsu* by a quantification theory from linguistic features of the *bunsetsu*. The most probable sequence of *bunsetsu* F0 shapes for a sentence are found in the F0 shape database by a probabilistic measure.

The probability that an F0 contour is observed for a sentence is defined by two kinds of probabilities, the F0 shape production and the F0 shape bigram. The latter is a probability of adjacent occurrence of two F0 shapes, like a word bigram in speech recognition. Several typical *bunsetsu* F0 shapes are extracted by clustering of training data and stored in the F0 shape database. The probability of the F0 shape production is computed for each *bunsetsu* based on the distribution of linguistic features in the cluster.

1. Introduction

The F0 generator is responsible for the naturalness and the intonational richness of synthetic speech that a TTS produces. Modeling the relationship between linguistic information and prosody is an important issue not only in speech synthesis, but also in speech recognition and dialogue processing[1]. This paper propose a stochastic F0 model which computes the probability of an F0 contour given a sentence, like a similar formulation used for speech recognition. It can generate an F0 contour for speech synthesis by selecting the most probable contour from possible ones. The stochastic F0 model could be also easily incorporated into speech recognition so as to score sentence hypotheses with a prosodic aspect, because most of the current speech recognition systems rely on the a probabilistic measure and select the most probable word sequence.

These days, many corpus-based approaches of F0 modeling have been proposed[2, 3, 4]. The corpus-based approaches including a stochastic modeling generally require many training data to estimate parameters or probabilities of the model. It is important that we can easily prepare many training data with F0 parameters, such as pitch values, phrase and accent commands[5], Tilt parameters[6], and so on, which describe F0 characteristics and are controlled by the model. The superpositional F0 models, such as Fujisaki's model[5], decompose a sentence F0 contour into two types of components of the major phrase and the minor phrase. The automatic decomposition of an F0 contour is not an easy task. Alternative description scheme of the F0 contour should be discussed for the corpus-based approach.

An F0 contour of a Japanese sentence consists of F0 patterns of the "prosodic word"[7], which forms an accent type and is often equal to *bunsetsu*. The *bunsetsu* is a syntactic unit of Japanese and it consists of a content word followed by function words. It is easy to identify *bunsetsus* in a sentence based on the morphological analysis. The prosodic words with the same accent type and the same syllable length fundamentally invoke similar F0 patterns although the context information or the focus deform the F0 pattern. Most of F0 models for spoken Japanese have a mechanism describing characteristics for the F0 pattern of the prosodic word, like Fujisaki's model[5]. In our proposed model, typical F0 pattern of the *bunsetsu* are extracted by clustering. A new idea is proposed that the probabilities of F0 pattern is computed based on the result of the clustering.

2. Method

The F0 contour of a sentence is represented by concatenation of F0 patterns of *bunsetsu*. An *bunsetsu* F0 pattern is described both by the *bunsetsu* F0 average which is an average pitch value in the *bunsetsu* and the *bunsetsu* F0 shape which is a normalized F0 contour of the *bunsetsu* with its average zero. Figure 1 shows the generation process of a sentence F0 contour. The F0 average and the F0 shape are independently predicted from linguistic features of the *bunsetsu* by the quantification theory and the stochastic model, respectively. Figure 2 shows a schematic F0 contour of a sentence based on the shape and the average of *bunsetsus* which are depicted by the solid and the dotted line, respectively. The *bunsetsu* F0 shapes are selected from the F0 shape database and biased by each F0 average.

2.1. *Bunsetsu* F0 average

The *bunsetsu* F0 average is predicted by the quantification theory (type I), which estimates a numerical value from several categorical input features using the equation[8]

$$\hat{y} = \bar{y} + \sum_{j=1}^{M} \sum_{v=1}^{V_j} \delta_{jv} x_{jv} \tag{1}$$

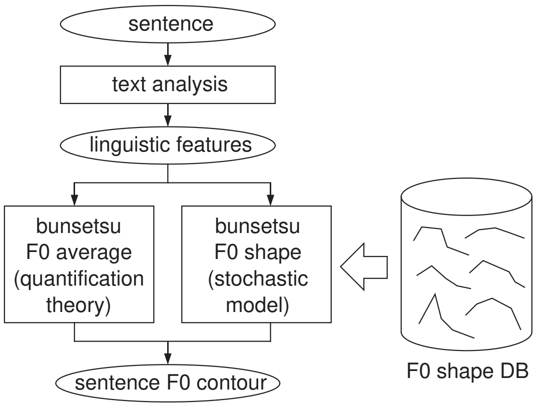


Figure 1: *Process of generation of sentence F0 contour.*

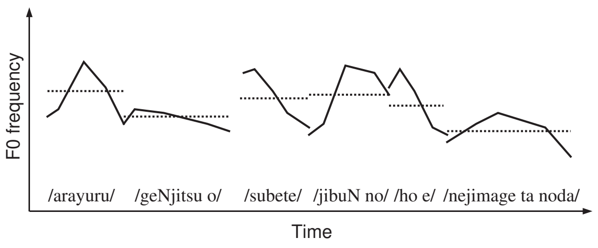


Figure 2: *F0 contour based on the shape and the average of bunsetsus.*

where $\hat{y}$ is the estimated value, $\bar{y}$ is the mean value of training samples, M is the number of input features, V_j is the number of value types for the j-th feature, and δ_{jv} is defined as

$$\delta_{jv} = \begin{cases} 1 & \text{if the } j\text{-th feature takes the } v\text{-th value} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The weights, x_{jv}, are obtained by minimizing the total RMS error of estimated values.

The basis of this *bunsetsu* F0 average modeling is the same as the global model in Abe's model[9]. The linguistic features of the *bunsetsu* used for the quantification theory are listed as follows.

- The lexical accent type of the preceding, the current, and the following *bunsetsu.* (3 categories)
- The part-of-speech of the preceding, the current, and the following *bunsetsu.* (8 categories)
- The syllable number in the preceding, the current, and the following *bunsetsu.* (7 categories)
- The preceding and the following boundary types of the *bunsetsu.* (4 categories)

The F0 average is independently predicted for each *bunsetsu.*

2.2. *Bunsetsu* F0 shape

A *bunsetsu* F0 shape is predicted by selecting a shape from the F0 shape database based on a probabilistic measure. Probability of a sequence of F0 shapes for the sentence is maximized. Let a sentence be composed of a sequence of *bunsetsus*, $B_1, B_2, \cdots, B_N(= B)$, where N is the number of *bunsetsus* of the sentence. The sequence of *bunsetsu* F0 shapes for the sentence, $x(1), x(2), \cdots, x(N)(= X)$, are obtained by maximizing

$$P(X|B) = \prod_{i}^{N} b_{x(i)}(B_i) a_{x(i-1)x(i)}{}^{r} \quad (3)$$

where $b_k(B_i)$ is a probability that B_i produces k-th F0 shape, and a_{kl} is a bigram probability that k-th and l-th F0 shape are adjacently occurred. This equation is similar to the formula used for continuous speech recognition, and $b_k(B_i)$ and a_{kl} correspond to the acoustic and the linguistic probability, respectively. The r is a weighting factor of the F0 shape bigram.

The probability of the F0 shape production, $b_k(B_i)$, is computed by

$$b_k(B_i) = \prod_{j}^{M} P_{kj}(f_j(B_i))^{q_j}. \quad (4)$$

The function $f_j(B_i)$ extracts a value of the j-th feature from a *bunsetsu* B_i. The $P_{kj}(f)$ is a probability that a value of the j-th feature is f in the k-th F0 shape. The q_j is a weighting factor of j-th feature. The feature set used in the equation (4) is the same as that for the F0 average prediction mentioned in 2.1. The q_j is 15.0 for three features of the current *bunsetsu*, the lexical accent type, the part-of-speech, the syllable number, and 1.0 for the others in this paper.

When the length of the selected F0 shape does not match with the target duration of the *bunsetsu*, the selected F0 shape is linearly lengthened or shortened.

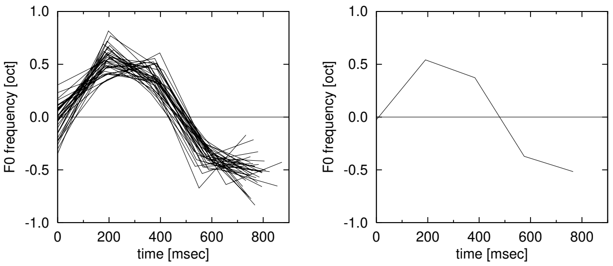
2.2.1. Clustering of F0 shapes

The F0 shape database is build by clustering of F0 shapes of the *bunsetsu.*

Each F0 pattern of the *bunsetsu* represented in the log scale is smoothed by the linear regression after the interpolation of unvoiced segments. The F0 pattern is approximated by fitting of connected four lines, that is five nodes, so that parametric representation facilitates computing the centroid of patterns in a cluster. The Japanese accent pattern of a word or a *bunsetsu* is described as a sequence of high and low tones with at most one accent nucleus. Four is the minimum number of lines which are capable of description of such characteristics. The approximation error is minimized by the steepest descent method. The parametric F0 pattern of the *bunsetsu* is unbiased so that the average is zero, and clustered by the LBG algorithm. The centroid in a cluster is computed by averaging coordinates of five nodes of the fitting lines.

Although the F0 shape is represented by four lines, the difference of F0 shapes is measured for interpolated time sequences of the F0 value. The distance between two F0 shapes, $p_1(t)$ and $p_2(t)$, is defined as

$$D = \frac{\sum_{t=0}^{min(T_1,T_2)} |p_1(t) - p_2(t)|}{min(T_1, T_2)} + w \times |T_1 - T_2| \quad (5)$$



(a) samples (b) a centroid

Figure 3: *The F0 shape samples and the centroid in a cluster.*

where T_i is the length of the F0 shape $p_i(t)$. The first and the second terms represent the difference of F0 values and the pattern length, respectively. The w is a weighting factor and it is 0.2 in this paper.

Figure 3 (a) and (b) shows an example of F0 shapes in a cluster and a centroid of the cluster, respectively.

2.2.2. Probability of F0 shape production

The probability of F0 shape production, $P_{kj}(f_j(B_i))$ in the equation (4), is computed based on the frequency distribution of values for the j-th feature in the k-th cluster. Let $C_{kj}(f_v)$ and $f_j(B_i)$ be the number of patterns which have the value f_v for the j-th feature in the k-th cluster, and a value of the j-th feature of B_i, respectively. The $P_{kj}(f_j(B_i))$ is defined as

$$P_{kj}(f_j(B_i)) = \frac{C_{kj}(f_j(B_i)) + \varepsilon}{\sum_{v=1}^{V_j}(C_{kj}(f_v) + \varepsilon)}. \tag{6}$$

To avoid zero probabilities due to the sparseness, all occurrence counts of values in the cluster are biased by ε for the flooring. The ε is 0.5 in this paper.

2.2.3. Probability of F0 shape bigram

For computing the probability of F0 shape bigram, a_{kl} in the equation (3), the bigram context includes a beginning of the sentence and a pause as well as typical F0 shapes. The probabilities are trained after the same flooring procedure as the probability of the F0 shape production mentioned in 2.2.2.

3. Evaluation

3.1. Speech data

The proposed model of F0 contour generation is trained on phonetically balanced 453 sentences uttered by a male professional speaker. These sentences include about 3100 *bunsetsus*. Another set of 50 sentences by the same speaker is used for the evaluation.

3.2. Four-line fitting of F0 shapes

The RMS error for approximation of F0 contour by the four line fitting is 0.089 [oct]. The error is computed for voiced segments of original speech. This evaluation manner is also used for following experiments.

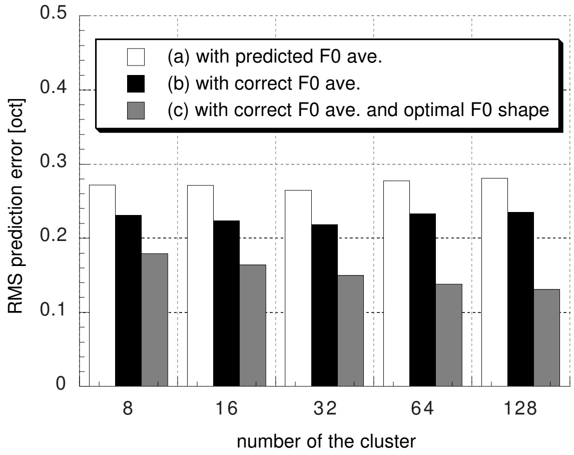


Figure 4: *The RMS errors for the number of the F0 shape clusters.*

3.3. Prediction of *bunsetsu* F0 average

The multiple correlation coefficient of the quantification theory for predicting the *bunsetsu* F0 average is 0.88. If the multiple correlation coefficient is 1.0, the prediction is perfect. The model of the *bunsetsu* F0 average has good performance. The RMS error for predicting the *bunsetsu* F0 average is 0.182, 0.171[oct] for the closed and the open data, respectively.

3.4. Prediction of *bunsetsu* F0 shape

3.4.1. The number of the clusters

The proposed method is evaluated for various numbers of the cluster in the F0 shape database. The F0 shape bigram, r in the equation(3), is 1.0 in this experiment. Figure 4 shows RMS errors of the sentence F0 contour for the open data, changing the number of the clusters, 8, 16, 32, 64, and 128. Figure 4 (a) is the prediction errors by the proposed method. In this experiment, 32-cluster model minimizes the errors by the proposed method. However, the optimal number of the clusters is dependent on the size of the training data, and it may be increased if more training data are available.

To analyze the prediction errors, F0 contours are generated under two more different conditions. Figure 4 (b) and (c) are the results of prediction with the correct *bunsetsu* F0 average. The condition (c) is the same as (b) except that optimal F0 shapes which minimize the prediction error of the F0 contour are selected. It indicates the lower bound of the prediction error when all F0 shapes are ideally selected from the database. Thus, the error is decreased as the number of the clusters is increased for (c). We can estimate the errors due to the incompleteness of the cluster selection from the difference between (b) and (c). If we can select the optimal cluster, the error is reduced by about 0.1 [oct] for 128 clusters. The difference between (a) and (b) indicates the errors caused by predicting *bunsetsu* F0 average.

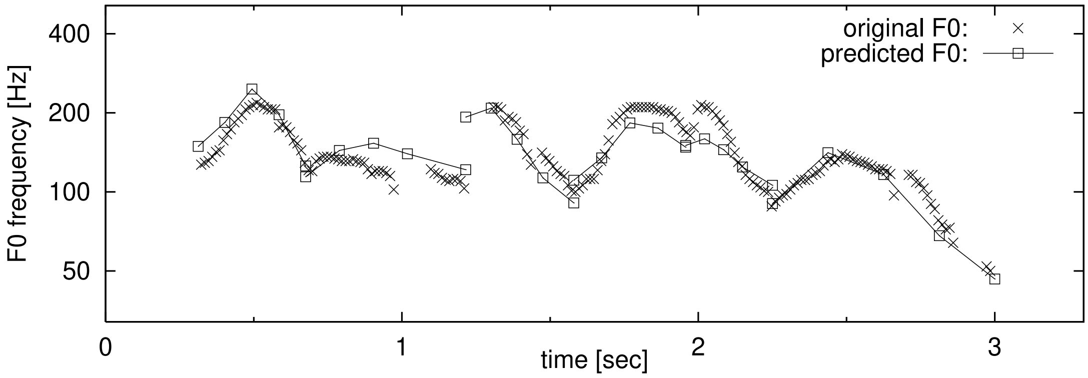


Figure 6: *Original and predicted F0 contours of a sentence.* (*/arayuru geNjitsuo subete jibuNno hoe nejimagetanoda/*)

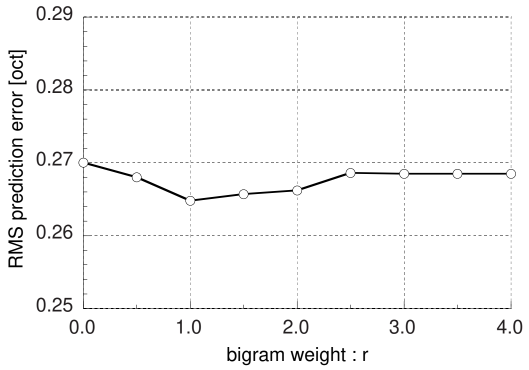


Figure 5: *The RMS errors for changing the weight of F0 shape bigram, r.*

3.4.2. The weight of the F0 shape bigram

We investigated the performance of the F0 contour model changing the weight of the F0 shape bigram, r in the equation (3). The RMS errors are shown in Figure 5. In this Figure, $r = 0$ means that the *bunsetsu* F0 shape is independently selected for each *bunsetsu*. The introduction of the F0 shape bigram decreases the prediction errors, and $r = 1$ gives the minimum prediction errors. However, it produces a small effect because the context information of the preceding and the following *bunsetsu* is embedded in the features of the *bunsetsu*.

Figure 6 shows an example of the F0 contour of a sentence predicted by the proposed model for a sentence. The predicted F0 contour is very similar to the original one.

4. Conclusions

A new stochastic F0 contour model is proposed. The F0 contour of a sentence is represented by concatenation of the F0 patterns of *bunsetsu*. An F0 pattern of the *bunsetsu* is composed of the F0 average and the F0 shape. The most probable sequence of *bunsetsu* F0 shapes in a sentence are found by a probabilistic measure in the F0 shape database. The RMS prediction errors of the F0 contour are 0.267[oct]. It was verified that this model has a good performance of predicting an F0 contour for the speech synthesis application, but several tunings, such as optimization of weights, increase of training data, and so on, are necessary to reduce the prediction errors.

This model probabilistically represents the relation between the F0 contour and linguistic features of a sentence. It can be easily applied to speech recognition applications, such as re-scoring of sentence hypotheses for N-best recognition result using prosodic information. Future works include improvement of the model and application to the speech recognition.

5. References

[1] "Computing Prosody", eds. Y. Sagisaka, N. Campbell, N. Higuchi, Springer (1997).

[2] F. Malfrère, T. Dutoit, and P. Mertens : "Automatic Prosody Generation Using Suprasegmental Unit Selection", Proc. of Third ESCA/COCOSDA Workshop on Speech Synthesis, pp.323-328 (1998).

[3] K. E. Dusterhoff, A. W. Black, and P. Taylor : "Using Decision Trees within the Tilt Intonation Model to Predict F0 Contours", Proc. of Eurospeech '99, 4, pp.1627-1630 (1999).

[4] M. Isogai and H. Mizuno : "A New F0 Contour Control Method Based on Vector Representation of F0 Contour", Proc. of Eurospeech '99, 2, pp.727-730 (1999).

[5] H. Fujisaki and K. Hirose : "Analysis of Voice Fundamental Frequency Contours for Declarative sentence of Japanese", J. Acoust. Soc. Jpn. (E), 5, 4, pp.233-242 (1984).

[6] P.Taylor : "The Tilt intonation model", Proc. of ICSLP '98 4, pp.1383-1386 (1998).

[7] H. Fujisaki, K. Hirose, and N. Takahashi : "Manifestation of Linguistic and Para-Linguistic Information in the Voice Fundamental Frequency Contours of Spoken Japanese", Proc. of ICSLP '90, 12-1, pp.485-488 (1990).

[8] C. Hayashi : "On the quantification of qualitative data from the mathematicostatistical point of view", Ann. Inst. Statist. Math., 2 (1950).

[9] M. Abe and H. Sato : "Two Stage F0 Control model Using Syllable Based F0 Units", Proc. of ICASSP '92, II, pp.53-56 (1992).