GENERATION OF F0 CONTOUR USING STOCHASTIC MAPPING AND VECTOR QUANTIZATION CONTROL PARAMETERS

Heo-Jin Byeon, Yeon-Jun Kim, Yung-Hwan Oh

Department of Computer Science
Korea Advanced Institute of Science and Technology
E-mail: bhjin@adam.kaist.ac.kr

ABSTRACT

This paper introduces an F0 contour generation method for text-to-speech synthesis using stochastic mapping and vector quantization control parameters. This model uses a new F0 contour labelling scheme based on the RFC (Rise/Fall/Connection) model [1], which describes F0 contour patterns with seven F0 labels and three pause labels. This paper also suggests an efficient selection method for control parameters instead of using the mean values of the control parameters. We achieved 78.06% accuracy in the F0 label prediction and 95.87% accuracy in the pause label prediction using this model. The experimental results shows that synthesized speech using vector quantization control parameters is more natural than using the mean values of the feature parameters.

1. INTRODUCTION

Most existing F0 contour generation systems are based on a rule-based model which generates definitive F0 contours with averaged values. And they usually take interpolation or filtering approaches[2]. Therefore, rule adjustments or rebuilding are required for spontaneous speech synthesis, not for read speech. And the F0 contour generated from interpolation or filtering is different from the original contour.

Another approach is a corpus-based model which trains HMM or neural networks automatically [3, 4] or which generates synthetic parameters using patterns extracted from natural speech waveforms of very large corpora [5]. But it is an inadequate way of applying the corpus-based model to text-to-speech synthesis since syntactic information is hardly involved in F0 contour generation, and very large corpora are needed.

The proposed approach attempts to build up a method that facilitates the extraction of F0 contour labels from the syntactic information of given sentences, using stochastic mapping instead of rules. Another goal is to model the details of natural intonation as closely as possible using the subdivided prosody labels.

2. MODEL OVERVIEW

The proposed model has a three-layered architecture, as shown in Figure 1, (1) the F0 label prediction layer, (2) the control parameter prediction layer, and (3) the F0 contour generation layer.

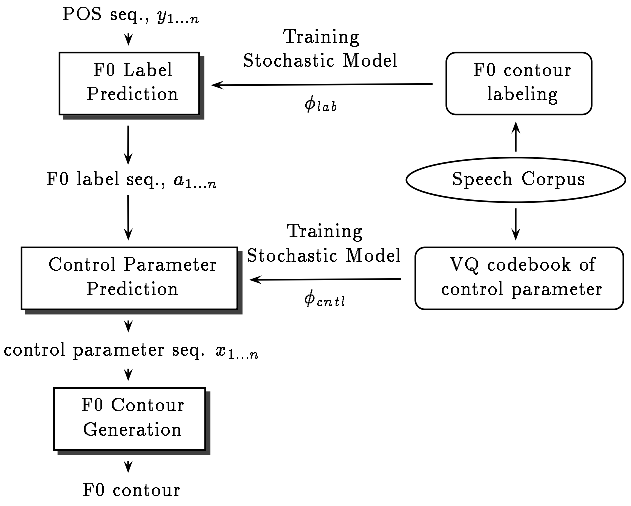

POS seq., $y_{1...n}$
F0 Label Prediction
Training Stochastic Model
ϕ_{lab}
F0 contour labeling
Speech Corpus
F0 label seq., $a_{1...n}$
Control Parameter Prediction
Training Stochastic Model
ϕ_{cntl}
VQ codebook of control parameter
control parameter seq. $x_{1...n}$
F0 Contour Generation
F0 contour

Figure 1: The Proposed F0 Contour Generation Model

The F0 label prediction layer generates F0 labels from the syntactic information of the input sentence using the stochastic mapping trained with hand-labeled F0 labels. The control parameter prediction layer selects the control parameters proper to the given F0 labels which are generated at the previous layer. We use stochastic mapping between the F0 labels and the control parameters, which are trained by the VQ codebook of the control parameters extracted from the speech corpus. At the F0 contour generation layer, the F0 contour is generated using the results of the control parameter prediction layer.

Instead of direct mapping from syntactic information(S) to control parameters(P), $(P(P|S))$, the insertion of an F0 label(L) prediction $(P(P|L)P(L|S))$ provides a more effective way of describing the F0 contours which have various shapes according to the F0 label sequence.

3. F0 LABEL PREDICTION

To describe F0 contour in detail, we define seven F0 labels which represent the fluctuation of the F0 contour and three prosodic boundary labels based on the pause duration, as shown in Table 1.

The F0 labels defined in this paper are modifications of *Taylor*'s work, *RFC* (rise/fall/connection) model [1]. Our F0 labels make it possible to assign each F0 label to a single morpheme only and reflect the contextual effect, for example, successive rise(s) and successive fall(g).

Table 1: F0 and Prosodic Boundary Labels

	label	meaning
F0 Label	r	rise
	s	successive rise
	f	fall
	g	successive fall
	p	peak
	v	valley
	c	connection
Boundary Label	.	no pause
	/	minor boundary
	%	major boundary

The input parameters of the F0 label prediction layer are extracted from the in-order sequence of the syntactic parse tree of the input sentence as shown in Fig. 2. Let the in-order sequence of the syntactic tree for the input sentence be S. We can also express syntactic parse tree S with eq. (1).

$$S = \alpha_0 \beta_1 \alpha_1 \beta_2 \cdots \alpha_{n-1} \beta_n \alpha_n \tag{1}$$

In eq. (1), α_0 represents a start symbol of the input sentence, α_n an end symbol, α_i a fired rule number of a given non-terminal node of the syntactic tree, and β_i a fired rule number of a given terminal node, respectively. We use $\alpha_{i-1}\beta_i\alpha_i$ for the ith input parameter of the stochastic model, s_i. The first element of the input parameter is equivalent to the last element of the previous input parameter. Such input parameter structure can represent the prior and the subsequent syntactic information of a given morpheme. When β_i represents a given morpheme in the ith input parameter, s_i, α_{i-1} represents the prior information of the syntactic structure of β_i and α_i the latter information. We use 96 rules for non-terminal node generation and 24 rules for terminal node generation.

The output of the F0 prediction layer is the F0 label sequence, which is matched to a given syntactic tree. Let the predicted F0 label sequence be L, as shown in eq. (2).

$$L = p_0 f_1 p_1 f_2 \cdots p_{n-1} f_n p_n \tag{2}$$

In eq. (2), p_0 represents a start symbol, p_n an end symbol, p_i a pause label, and f_i an F0 label. The ith output parameter of stochastic model, l_i, corresponding to the ith input parameter, s_i, is $p_{i-1}f_ip_i$. Such output parameter structure represents the F0 contour including both the prior and the subsequent information. Though there are 112 possible output parameters, we used only 38 output parameters which appeared in the speech corpus.

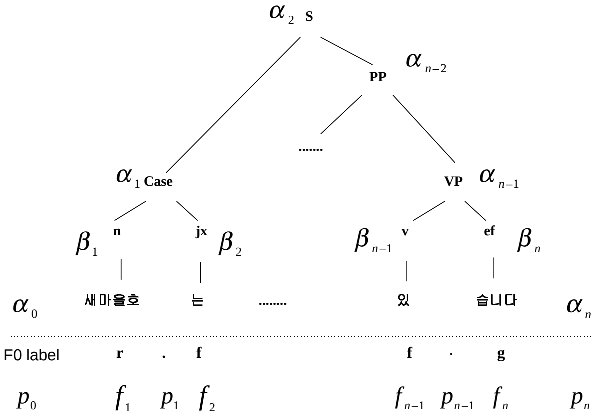


Figure 2: Syntactic Parse Tree, (α_i and β_i represent syntactic information and f_i and p_i F0 and pause label.)

In the F0 label prediction layer, we choose the most probable sequence of F0 contour labels($l_{1 \dots n}$) for a given piece of syntactic information ($s_{1 \dots n}$) of a sentence using the **stochastic mapping** function as in eq.(3). Applying *Bayes decision rule* and the 1st-order *Markov assumption*, the stochastic mapping function can be approximated with the multiplication of output probabilities and transition probabilities(eq.(6)).

$$\phi_{lab}(s_{1 \dots n}) \quad = \quad \arg\max_{l_{1 \cdots n}} P(l_{1 \dots n} | s_{1 \dots n}) \tag{3}$$

$$= \arg\max_{l_{1\cdots n}} \frac{P(s_{1\cdots n}|l_{1\cdots n})P(l_{1\cdots n})}{P(s_{1\cdots n})} \quad (4)$$

$$= \arg\max_{l_{1\cdots n}} P(s_{1\cdots n}|l_{1\cdots n})P(l_{1\cdots n}) \quad (5)$$

$$\cong \arg\max_{l_{1\cdots n}} \prod_{i=1}^{n} P(s_i|l_i)P(l_i|l_{i-1}) \quad (6)$$

The use of the stochastic model has the advantage of being able to find the relations between given syntactic information and F0 contour label because the mapping problems can be solved automatically with the *Viterbi* search algorithm.

4. CONTROL PARAMETER PREDICTION

To generate F0 contours with F0 labels, four control parameters are adopted :
$\{tilt,\ gradient_{1st-half},\ gradient_{2nd-half},\ start\}$.
Figure 3 shows the meaning of each element of the control parameter. *tilt* represents the shape(rise(+)/fall(-)) and the bending position in a segment of F0 contour. $gradient_{1st-half}$ and $gradient_{2nd-half}$ refer to the gradients of the 1st- and 2nd-half of the segment. *start* represents the beginning frequency(Hz) of the F0 contour.

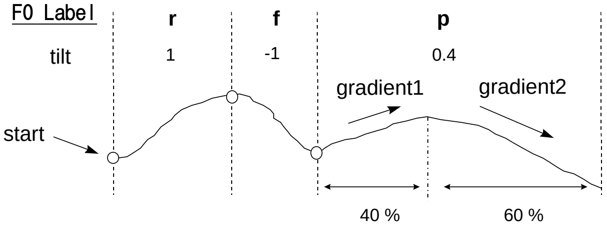


Figure 3: Control Parameters for F0 Contour Generation

Another distinguishing feature in our work is the adoption of **vector quantization (VQ) control parameters** instead of several F0 patterns or averaged values. In general, the mean of the control parameters or several F0 patterns are used to generate the F0 contour. But they are not enough to represent all the various F0 contours. The VQ control parameters provide a richer and more comfortable scheme generating detailed and close F0 contours. With the *K-means* algorithm, the control parameters are clustered independently of the F0 labels to reduce the intra-variance of each cluster. In this work, we used 130 clusters, of which standard variations are less than 20% from the centroid of each cluster.

In the control parameter prediction layer, given the sequence of the F0 label(l_i), we can get the best codeword sequence of the control parameter vectors (p_i) by using the stochastic mapping model (eq.(8)) in the same way as the F0 label prediction layer.

$$\phi_{cntl}(l_{1\cdots n}) = \arg\max_{p_{1\cdots n}} P(p_{1\cdots n}|l_{1\cdots n}) \quad (7)$$

$$\cong \arg\max_{p_{1\cdots n}} \prod_{i=1}^{n} P(l_i|p_i)P(p_i|p_{i-1}) \quad (8)$$

5. F0 CONTOUR GENERATION

Finally, the F0 contour of a sentence is derived from eq.(9) and eq.(10) [6].

$$f_0 = \begin{cases} A - 2A(\frac{t}{D})^2, & 0 \le t < \frac{D}{2} \\ 2A(1-\frac{t}{D})^2, & \frac{D}{2} \le t < D \end{cases} \quad (9)$$

$$F_0 = start - A - f_0 \quad (10)$$

In eq.(9), A is the amplitude of F0 frequency variation of a segment and D is the duration factor. A is the result of the multiplication of $gradient_{1st-/2nd-half}$ and D. We take the last F0 frequency of the previous segment as *start* when a given segment follows a previous one, otherwise *start* itself is used.

6. EXPERIMENTAL RESULTS

The proposed model was trained with 158 sentences spoken by a female announcer and tested with 19 sentences. The training speech corpus has hand-labeled 2037 F0 labels and 1879 pause labels, and the test corpus has 237 F0 labels and 218 pause labels, respectively.

At first, we evaluated the efficiency of the F0 label prediction layer with the prediction rate which is the correction rate of the generated F0 labels compared with the hand-labeled F0 labels. The prediction rate of the F0 label prediction layer is shown in Table 2. As shown in Table 2, we achieved an 88.6% prediction rate in the training sentences and 85.5% in the test sentences at the F0 label prediction layer.

Table 2: The Result of F0 label Generation

	training data	test data
F0 label	81.10%	78.06%
pause label	96.75%	95.87%
total	88.61%	86.59%

To evaluate the effectiveness of the control parameter prediction layer, we carried out two subjective tests, an MOS and a preference test, which compare synthesized speech using VQ control parameters with synthesized speech using averaged control parameters.

Table 3 shows the result of the MOS test which 20 native listeners participated in. In the MOS test, synthesized speech using VQ control parameters got a higher score than it did when using averaged control parameters.

Table 3: The Result of MOS Test

	MOS	S.D
averaged control parameter	3.6	1.02
VQ control parameter	3.9	0.69

Table 4: The Result of Preference Test

averaged control para.	VQ control para.	not distinguished
22 %	55 %	23 %

The result of a preference test on 20 native listeners reveals a preference for synthesized speech using VQ control parameters over speech using averaged parameters, as shown in Table 4. In the preference test, there is little preference difference between the two synthesized speech when the sentence is short (less than 10 morphemes in a sentence). But when the sentence is long (more than 10 morphemes in a sentence), listeners prefer synthesized speech using VQ control parameters to speech using averaged parameter values. Figure 4 shows the F0 contour generated by the proposed model.

7. CONCLUSION

In this paper, we proposed an F0 contour generation method for text-to-speech synthesis. The use of stochastic mapping and VQ control parameters reveals an easier and more elegant way of generating detailed and close F0 contours.

In this work, we used only the syntactic information of the input sentence to generate an F0 contour without considering the pitch pattern of the phoneme itself. To obtain more natural synthesized speech, the pitch pattern of the phoneme should be involved. In the future, we are going to improve our work by taking account of this factor.

8. REFERENCES

[1] Paul Taylor, "The rise/fall/connection model of intonation," Speech Communication, Vol.15, pp.169-186, 1994

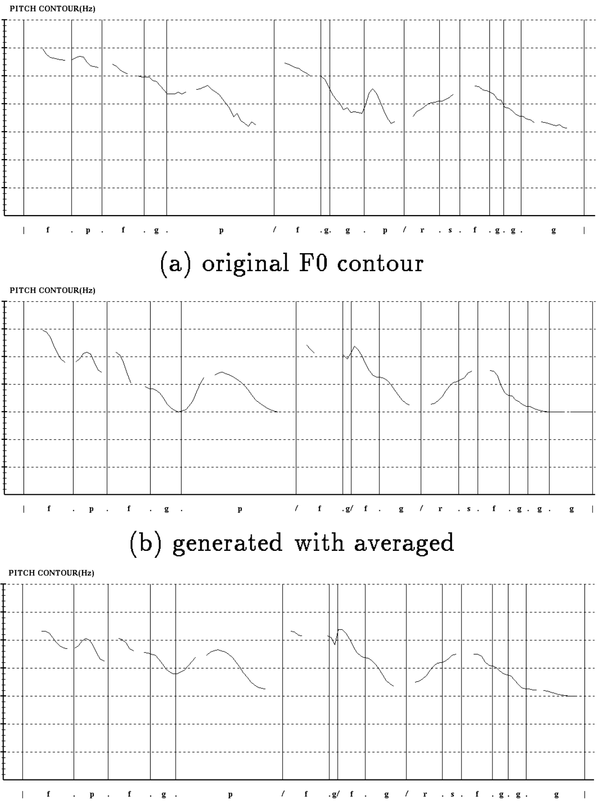
[2] H. Fujisaki, H. Kawai, "Realization of Linguistic Information in the Voice Fundamental Frequency Contour," *ICASSP'88*, pp.663-666, 1988

[3] Toshiaki Fukada, Yasuhiro Komori, Takashi Aso, Yasunori Ohora, "A Study on Pitch Pattern Generation using HMM-based Statistical Information," *Proc. ICSLP'94*, pp.723-726, 1994

[4] Michael S. Scordilis, John N. Gowdy, "Neural Network Based Generation of Fundamental Frequency Contours," *ICASSP'89*, pp.219-222, 1989

[5] Naohiro Sakurai, Takemi Mochida, Tetsunori Kobayashi, Katsuhiko Shirai, "Generation of Prosody in Speech Synthesis using Large Speech Data-Base," *Proc. ICSLP'94*, pp.747-750, 1994

[6] Paul Taylor, Alan W. Black, "Synthesizing Conversational Intonation from a Linguistically Rich Input," *Proc. of Second ESCA/IEEE Workshop on Speech Synthesis*, pp.175-178, 1994



(a) original F0 contour

(b) generated with averaged

(c) generated with VQ

Figure 4: F0 contour generated by the proposed model (prediction rate of F0 labels : 89.6 %), "서울발 대구행 열차표는 만이천원입니다." (The price of a train ticket from *Seoul* to *Daegu* is 12,000 won.)