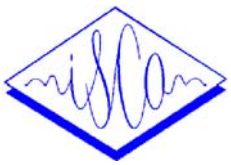


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A NEW F0 CONTOUR CONTROL METHOD BASED ON VECTOR REPRESENTATION OF F0 CONTOUR

Mitsuaki ISOGAI and Hideyuki MIZUNO
NTT Cyber Space Labs.
1-1 Hikari-no-oka Yokosuka-Shi Kanagawa 239-0847 Japan
isogai@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a new fundamental frequency(F0) contour control method based on vector representation of F0 contour. The main points of the proposed method are as follows;

(1) Desired F0 contours are created by selecting or modifying natural F0 contours held in a speech database.
(2) F0 contour selection is based on statistical estimation using a vector representation of F0 contour
(3) The selected F0 contour is modified to match the target context according to rules produced by statistical learning.

An evaluation by listening tests confirms the superior performance of our proposed over the conventional method approach to F0 modeling.

Keywords: Text-To-Speech, F0 contour control, speech database, vector presentation, statistical estimation

1. INTRODUCTION

To synthesize various types of speech is very important to extend the application area of synthesized speech. Especially for multimedia contents such as computer games, computer assisted instruction systems, and WWW pages, output speech that is more accurate, spontaneous, and emotional is required. In terms of speech type, fundamental frequency (F0) contour is one of the most important factors. It has already reported that F0 contour determines speaking style[1] and speaker[2]. Although speech synthesis has improved with recent advances in TTS[3][4], the result is still monotonous. Recently, the approach of corpus-based concatenative speech synthesis has been reported[5]. Also in the prosody research field, a corpus-base prosodic generation algorithm[6] has been proposed. These approaches make good use of natural speech samples held in a speech database to reproduce the desired prosodic characteristics. However, even if a large speech database is used, optimal speech can not be created because the desired sample is often not in the database.

This paper proposes a new approach to generating F0 contours. It consists of two ideas: utilization of original F0 contour if available, and selection and modification of the nearest original contour to the target.

The following section outlines of the proposed method. It describes the F0 contour selection method. Because the database will not cover a sufficient number of patterns, F0 contour generation where contour selection is based on statistical estimation is described. Finally, Section 3 evaluates the proposed method.

2. NEW F0 CONTROL METHOD

2.1 Outline

The block diagram of F0 contour control is shown in Figure 1. The F0 contour generation process is divided into two stages as follows;

(1) If there is an original F0 contour in the speech database that is sufficient given the target context, it is selected and used without change as the F0 contour for the target context.

(2) In all other cases, the nearest F0 contour is selected and modified accordingly. Contour selection is based on statistical estimation with vector representation of F0 contour.

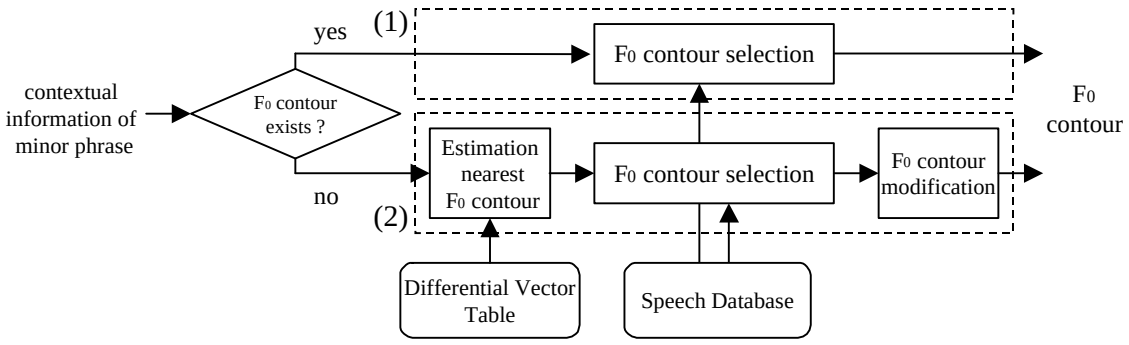


Figure 1: Block diagram of F0 contour control

2.2 Locating Natural F0 Contour

All the F0 contours are classified based on contextual information. We define those classes as the F0 contour group. Details of the factors are explained in section 3. The input text is divided to minor phrases. Each minor phrase is classified into an F0 contour group according to its contextual information. A natural F0 contour, which is the centroid of the F0 contour group, is selected from a speech database.

2.3 Control of F0 Contour Based on Statistical Estimation

2.3.1 Difinition of Distance between each F0 Contour Groups

The following procedures are performed off-line. Because it is difficult to estimate and modify continuous F0 contours, all F0 contours in the speech database are represented using vectors. Figure 2 shows a vector representation of an F0 contour. We define the vector of F0 contour **A** as

$$\mathbf{A} = (h_1 - h_0, \ldots, h_{m-1} - h_{m-2}, \underbrace{0, \ldots, 0}_{M-m}, h_m - h_{m-1}, h_{m+1} - h_m)$$

where $\{ h_i : 1<i<m\}$ is the F0 value at the center of the vowel within the *i*-th syllable, h_0 is the F0 value at the top of the phrase, h_{m+1} is the F0 value at the end of the phrase, *m* is the number of syllables in the phrase, and *M* is the maximum number of syllables in the speech database. The vector is normalized to yield a dimensionless term. The mean of all vectors in each F0 contour group is defined as the typical vector of the F0 contour group. The difference in F0 contours between each F0 contour group is defined as the differential vector of the typical vector of each F0 contour group, i.e. the norm of the differential vector between each vector of F0 contours means the distance between the F0 contour groups.

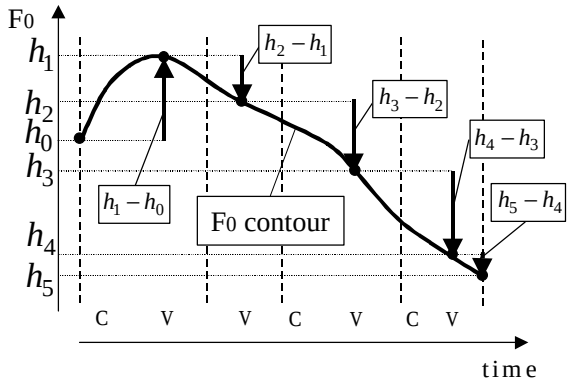
2.3.2 Differential Vector Estimation

Following procedures in this paragraph are performed off-line. Qualification theory (type one)[7], is applied to estimate an element of the differential vectors from the F0 contour group in the speech database to an F0 contour group not in the database. This yields a type of factor analysis, and can formulate the relationship between categorical and numerical values using the relationship

$$\hat{y}_i = \tilde{y} + \sum_f \sum_c a_{fc} \delta_{fc}(i)$$

where $\hat{y}_i$ is the estimated value of the *i*-th sample, $\tilde{y}$ is the mean of all samples, and $\delta_{fc}(i)$ is the characteristic function:

$$\delta_{fc}(i) = 1 : \text{if the } i\text{-th sample falls into category } c \text{ of factor } f,$$
$$0 : \text{otherwise.}$$



vector $\mathbf{A} = (h_1 - h_0, h_2 - h_1, h_3 - h_2, 0,0,0,0, h_4 - h_3, h_5 - h_4)$

Figure 2: Vector representation of F0 contour

a_{fc} is obtained by minimizing the estimation error

$$\sum_i (\hat{y}_i - y_i)^2$$

The factor $\{f\}$ and the category $\{c\}$ represent the contextual information of the F0 contour group not in the speech database. Details of factors of contextual information and its categories are explained in section 3. In this estimation, the number of generated models is the number of F0 contour groups in the speech database multiplied by *M*+1. All a_{fc} and $\tilde{y}$ are stored the table of differential vector.

2.3.3 Selection of Proper F0 Contour

The following procedures are performed on-line. The F0 contour that most suits the input phrase is selected by locating the minimum norm of the differential vector. The search method is as follows:

Step 1. All differential vectors from the F0 contour group in the speech database to the F0 contour group of the input phrase are generated using a table of differential vectors and contextual information of the latter F0 contour group.

Step 2. The F0 contour group that has the minimum norm of differential vector is selected as the nearest F0 contour group.

2.3.4 Modification of F0 Contour

The proper F0 contour in the speech database is extracted. It is modified to match the contextual information of the input phrase; the values are generated by linearly interpolating the differential vector at each frame.

3. EVALUATION

3.1 Conditions of Evaluation

3.1.1 Speech Database

A speech database that contains 503 phonetically balanced sentences (total of 3365 minor phrases) was used in the evaluation test. The database consisted of manually segmented phoneme boundaries, prosodic and linguistic information, and manually corrected F0 contours[8].

Table 4: Results of subjective evaluation
(a) Proposed method, (b) Proposed method,
(b) Proposed method, (d) Conventional method,
and (e) Original

methods	(a)	(b)	(c)	(d)	(e)
score	3.28	3.32	3.16	3.00	3.98
RMS error (Hz)	16.62	18.26	18.66	49.08	

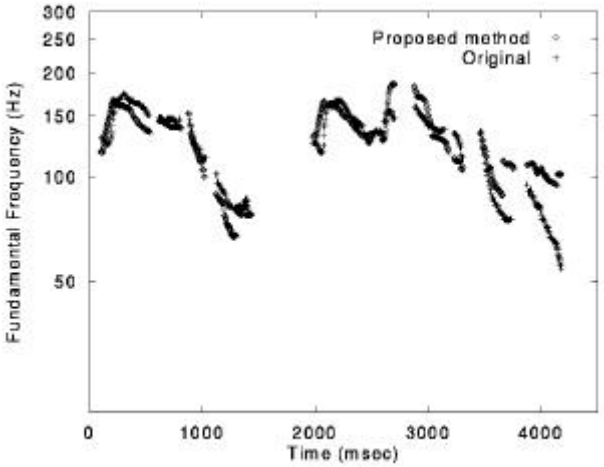


Figure 3: Example of F0 contours
Proposed method(a) and original(e)

base.

The results are shown in Table 2, the difference in RMS error was about 0.75Hz. These results show the accuracy of the estimation method.

Table 3 shows the averages of the partial and multiple correlation coefficients of each differential vector model. In terms of partial correlation coefficients, the results show that "the number of syllables in current phrase" and "accent type of current phrase" are the most important factors. In terms of "accent type of preceding phrase", the 1st model is more important than the others. That is, the 1st syllable was affected strongly by the accent type of the preceding phrase. "accent type of following phrase" has small influence. In terms of "preceding boundary type", 1st and 2nd models dominate the others. This shows that 1st and 2nd syllables were affected strongly by the preceding phrase.

tegories of a minor phrase

	boundary type	
following phrase	preceding phrase	following phrase
one, other than one	tight connection, loose connection, top of a sentence, preceded by pause	tight connection, loose connection, end of a sentence, followed by pause

Table 2: RMS error from objective evaluation

data set	RMS error (Hz)
closed	8.08
open	8.83

3.3 Subjective Evaluation

Listening tests were carried out to evaluate F0 contour generation. Five sentences were synthesized with F0 contours generated in five different ways.

(a) Proposed method: F0 contours were generated using speech database that included all groups of F0 contours needed for synthesized text.

Table 3: Average of partial and multiple correlation coefficients by objective evaluation

	partial correlation coefficients						
		accent type			boundary type		
		preceding phrase	current phrase	following phrase	preceding phrase	following phrase	
1st	0.550	0.212	0.196	0.015	0.409	0.161	0.628
2nd	0.650	0.088	0.489	0.037	0.391	0.347	0.759
3rd	0.598	0.086	0.523	0.096	0.179	0.192	0.690
4th	0.456	0.060	0.620	0.017	0.142	0.120	0.675
5th	0.449	0.055	0.605	0.011	0.157	0.081	0.695
6th	0.430	0.016	0.645	0.016	0.103	0.096	0.731
7th	0.320	0.132	0.700	0.070	0.088	0.123	0.757
8th	0.393	0.034	0.475	0.008	0.291	0.191	0.619
9th	0.100	0.056	0.319	0.029	0.182	0.340	0.475

(b) Proposed method: F_0 contours were generated by modifying the content of the speech database that did not include any of the F_0 contour groups needed.

(c) Proposed method: Same as (b), but F_0 contours were not modified.

(d) Conventional method: An accentual component is generated by linearly interpolating average F_0 of syllables, then superposed on a phrase component to generate overall F_0 contours[3].

(e) Original speech: F_0 contours of original natural speech were used.

Methods (b) and (c) were designed to evaluate the estimation of differential vectors and the modification of F_0 contours.

Method (d) was intended to compare the proposed method to the conventional method. Method (e) was designed to compare the proposed method to natural speech. All F_0 contours were adjusted to match that of the original speech. Twenty-five sentences (5 sentences by the 5 methods) were presented to ten listeners through headphones.

Table 4 shows the results of the listening test. It shows the scores of each method, and the RMS error against original F_0 contours. It shows that the proposed methods(a)(b)(c) are better than conventional method(d). Because the estimation of differential vectors is accurate, (b) has about the same score as (a). Because F_0 contour modification is effective, (b) has higher score than (c) and the RMS error decreased.

Figure 3 shows an example of the F_0 contours generated by the proposed method (a) and original (e). It shows that the proposed method can generate F_0 contours that well match the original ones.

CONCLUSION

We proposed an F_0 contour control method based on vector representation of F_0 contours. It uses natural F_0 contours held in a speech database. If the speech database does not cover the input text, it selects the closest F_0 contour by statistical estimation and modifies it to match the text. Listening tests showed that the proposed method generates better quality F_0 contours than the conventional method. As a future work, we will apply this method to an actual TTS system and evaluate it. We also intend to apply this approach to various types of speech.

ACKNOWLEDGEMENT

We are grateful to the members of the Media Processing Project for their helpful discussions. We also thank Mr.Yamamori, the executive manager, for his continuous support of this work.

REFERENCES

[1] C.Sorin, D.Larreur, R.Llorca, “A rhythm-based prosodic parser for text-to-speech systems in French,” Proc. of 11th ICphs, Tallinn, vol.1 pp.125-128, 1987

[2] A.I.C.Monaghan, D.R.Lass, “Manipulating synthetic intonation for speaker characterization,” ICASSP’91,pp 453-456, 1991

[3] K.Hakoda, H.Sato, “Prosodic rules in connected speech sysnthesis,” Trans. IEICE Vol.63 D No.9, pp.715-722, 1980 (in Japanese)

[4] M.Abe, H.Sato, “Two-stage F_0 control model using syllable based F_0 units,” Proc. ICASSP’92, pp.118-121, 1992

[5] N.Campbell, “Prosody and the selection of units for concatenation synthesis,” Proc. of the second ESCA/IEEE workshop on shpeech synthesis, pp.61-64, 1994

[6] F.Malfère, T.Dutoit, P.Mertens, “Automatic prosody generation using suprasegmental unit selection,” Proc.of the third ESCA/IEEE workshop on speech synthesis, pp.323-328, 1998

[7] C.Hayashi, “On the quantification of qualitative data from the mathematicostatistical point of view,” Ann. Inst. Statist. 1950

[8] M.Abe, Y.Sagisaka, H.Kuwabara, “Fundamental frequency database with linguistic phonetic information,” JASA, Vol.86 Suppl.1 O8 pp.S36, 1989