최소 자승오차 방식을 이용한 세그먼트 피치패턴의 정형화

이 정철, 김 상훈
한국전자통신연구소, 자동통역연구실

A New Stylization Method using Least-Square Error Minimization on Segmental Pitch Contour

Jung-Chul Lee, Sang-Hun Kim
Automatic Interpretation Section, ETRI

Abstract

In this paper, we describe the features of the fundamental frequency ($F0$) contour of Korean read speech, and propose a new stylization method to characterize the $F0$ pattern of segments. Our algorithm consists of three stylization processes: the segment level, the syllable level, and the word level. For stylization of $F0$ contour in the segment level, we applied least square (LS) error minimization method to determine $F0$ values at initial, medial, and final position in a segment. In the syllable level, we stylized the $F0$ pattern of a syllable using the mean $F0$ values of LS modeled segments in a syllable. Finally, we determine the stylized $F0$ pattern of word with the mean $F0$ values of syllables in a word. With the mean $F0$ value of each word and style information for each word, syllable and segment, we reconstruct $F0$ contour of sentences. The simulation results show that the error is less than 10% of the actual $F0$ contour for each sentence. In perception test, there is little difference between the synthesized speech with the original $F0$ contour and the synthesized speech with the stylized $F0$ contour.

1 Introduction

Intonation pattern in speech has attracted attention from a number of points of view. Many researchers have devoted efforts in characterizing what different intonation patterns exist, what sorts of meaning are conveyed by these patterns, and how the temporal relation between $F0$ contour and the speech segments is governed by stress and syntax. According to these works, intonation plays an important role in the intelligibility and naturalness of speech [4]. And an adequate intonation synthesis program is an important prerequisite to any speech synthesizer with practical applicability for extended synthetic speech [5] [6].

Recently, many researches have been working in intonation modeling based on stylization method to yield natural synthetic speech [1] [2] [3]. Most stylization methods use the pitch values simply chosen at initial, medial (or maximum) and final position of each segment [3]. Hence, it is not considered as a good model of the segmental intonation contour. Therefore, we propose a new modeling method using least-square error. This paper is organized as follows: Section 2 describes the prosodic database and some features of $F0$ pattern in Korean. Section 3 explains the stylization model and the analysis method. Section 4 shows simulation results, and section 5 remarks conclusion.

2 Features in pitch variation

The speech material for present study consists of recordings produced by reading individual sentences of written text in isolation. The written text is composed of 16 sentences, 99 words, 212 syllables, and 630 segments and read by a female announcer. The recorded material was digitized at 16 KHz with 16 bit precision. The fundamental frequencies were extracted by ESPS S/W, while the markers for segments were labeled manually. Figure 1 shows the distribution of pitch deviation from each mean $F0$ of segments.

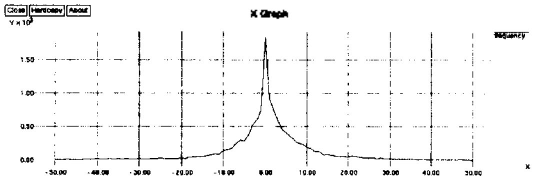

Figure 1: Distribution of pitch deviation in segment

The distribution of deviation is almost symmetric, and has high density around its mean. Table 1 presents cumulative distribution for $|x_i - \bar{x}| \le k$ in segments where x_i is the pitch value and $\bar{x}$ is the mean of each segment.

Table 1: Cumulative distribution(CD) for $|x_i - \bar{x}| \le k$ in segments

k	1	3	7	10	14	19	26	50
CD(%)	28.3	48.5	70.6	80.	86.8	91.4	95.2	100

The Cumulative distribution for segment shows that pitch in a segment varies in a limited range (k=50). Also, we calculated mean absolute deviation(MAD:$\frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$) for the segmental pitch contour and the result was 7.68 Hz. We can see that more than 90% of pitch variation is covered with k equal to 3*MAD. From the above results, we used MAD as a quantization step size in the stylization of pitch patterns.

Table 2 presents cumulative distribution for $|x_i - \bar{x}| \le k$ in segments where x_i is the pitch value and $\bar{x}$ is the mean of each syllable.

Table 2: Cumulative distribution(CD) for $|x_i - \bar{x}| \le k$ in syllables

k	1	3	7	10	14	19	24	69
CD(%)	19.4	35.4	57.2	69.	79.2	85.8	90.7	100

From Table 2, we can see that the pitch variance of syllables is higher than that of segments and this phenomenon is natural because a syllable consists of more than 1 segments. MAD for syllables is 10.1 Hz. The pitch variance for words is much higher than that of segments and MAD is 12.1 Hz.

3 Analysis and stylization of pitch pattern

3.1 Stylization in segmental level

To determine the style of segmental $F0$ pattern, $F0$ values simply chosen at three positions were coded in conventional methods. But, the estimate of pitch for speech signal usually has fluctuation on pitch contour due to estimation error or unstable vibration of vocal folds. Hence it is important to minimize these effects in the stylization of pitch pattern. To minimize these defects, we employed least square error minimization method to estimate the $F0$ values at three positions. First, the segmental pitch contour is piecewise linearly approximated by two straight lines, as shown in Figure 2.

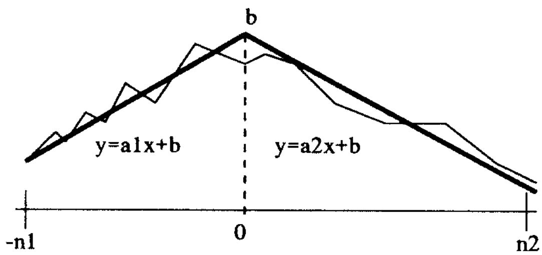

Figure 2: Two straight line approximation

Then, the sum of squared error between the original pitch contour and the linearly approximated pattern is

$$E = \sum_{i=-n1}^{0} (y_i - a_1 x_i - b)^2 + \sum_{j=1}^{n2} (y_j - a_2 x_j - b)^2. \qquad (1)$$

By setting the derivatives of E with respect to a_1, a_2, and b to zero,

$$\frac{\partial E}{\partial a_1} = -\sum_{i=-n1}^{0} (y_i - a_1 x_i - b) x_i = 0 \qquad (2)$$

$$\frac{\partial E}{\partial a_2} = -\sum_{j=1}^{n2} (y_j - a_2 x_j - b) x_j = 0 \qquad (3)$$

$$\frac{\partial E}{\partial b} = -\sum_{i=-n1}^{0} (y_i - a_1 x_i - b) - \sum_{j=1}^{n2} (y_j - a_2 x_j - b) = 0. \qquad (4)$$

Solving the polynomial equations (2),(3),(4), we calculate new pitch values and the mean $F0$ of the segment. Also, the position of peak or valley can be estimated if we find minimum of LS errors by varying n_1 and n_2. So this method can automatically detect pitch movement timing in a segment. Finally, new pitch values chosen at $-n_1, 0$, and n_2 are coded by symbols as shown in Figure 3.

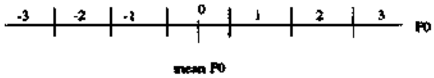

Figure 3: Interval distribution

Therefore, each segment can be stylized with 3 codes, e.g., 333, 630.

3.2 Stylization in syllable level

In this stage, the stylization of syllabic pitch pattern is carried out using the mean values of segments which compose a syllable. Using the mean values of segments, we reduced the dynamic range of $F0$ variation in a syllable. Korean syllables consist of V, CV, VC, CVC and each syllable is stylized when it is composed of more than two voiced segments. Firstly, we estimate mean $F0$ in a syllable, and then two or three mean $F0$ of segments are coded like in the previous stage 1.

3.3 Stylization in word level

Fortunately, since Korean words have no pitch accent, we can simplify the word $F0$ pattern like the segmental $F0$ contour. In this stage, the stylization of word pitch pattern is carried out using the mean values of syllables obtained in stage 2. The word pitch pattern is stylized when it is composed of more than two syllables. We find mean $F0$ of a word and code mean $F0$ of initial, middle, and final syllable like syllabic stylization.

3.4 Reconstruction of $F0$ contour

At this time, we have mean $F0$ of each word and the style information for each word, syllable and segment. The reconstruction of $F0$ contour starts from the computing the mean $F0$ value for each syllable using the style, mean $F0$ of a word. Then, we can find mean $F0$ of each segment by the same process. Finally, we reconstruct segmental $F0$ contour using the style, mean $F0$ and the duration information of each segment.

4 Simulation results

We performed three experiments. First, we experimented the robustness of our algorithm to the noisy pitch contour of a segment shown in Figure 4. Conventional stylization method with $F0$ values simply chosen at three positions resulted in 15.5Hz for MAD, but our algorithm using LS method yields 3.9Hz for MAD. This means our algorithm is robust to noisy pitch contour and makes reliable results.

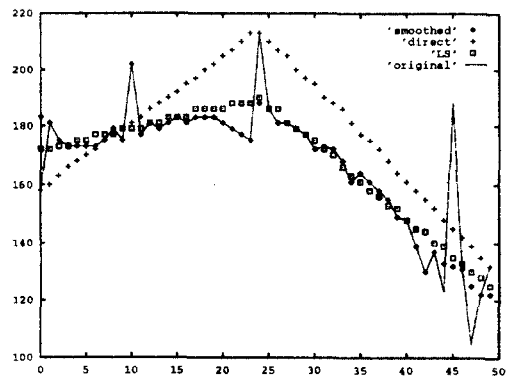

Figure 4: Comparison between original, smoothed, conventionally stylized, and LS stylized

Secondly, we found a number of stylized patterns in segment, syllable, and word. The possible number of stylized pattern is 347 when 7 level quantizer used. Practically, only 84 for segment, 19 for syllable, 53 for word were found in the speech material.

Table 3: Dominant style in segmental pitch pattern and its frequency of occurrence

style	freq.	style	freq.	style	freq.	style	freq.
333	110	342	22	243	12	620	8
432	41	531	20	233	11	431	8
433	38	441	16	324	10	144	8
234	28	630	15	621	10	343	8
334	25	640	14	424	8	450	8
532	24	622	12	522	8	332	7

Table 3 shows the dominant pitch pattern of segment and its frequency of occurrence. The portion of steady state style (333) is about 17%, and 41% is the portion of the style which has 3 in the middle position and 2, 3 or 4 in the first, or the third position. There exists 6.3%

occurrence of abrupt pitch fall like 036, while abrupt pitch rise is 17% of occurrence.

Table 4 and Table 5 show the dominant pitch pattern and its frequency of occurrence in syllable and word each, respectively.

Table 4: Dominant style in syllabic pitch pattern and its frequency of occurrence

style	freq.	style	freq.	style	freq.	style	freq.
33	62	60	20	360	3	433	2
42	55	51	10	560	3	460	2
24	21	342	3	333	2		

Table 5: Dominant style in word pitch pattern and its frequency of occurrence

style	freq.	style	freq.	style	freq.	style	freq.
423	8	333	5	414	4	33	4
234	5	60	5	342	4	51	4

We compared the reconstructed *F0* pattern with the original contour. It is shown in Figure 5. *F0* patterns were calculated by two linear functions between segment boundaries. MAD between the original contour and the reconstructed *F0* pattern was 6.24Hz. As a results, the reconstructed *F0* pattern appling LS method traces the original *F0* contour very well.

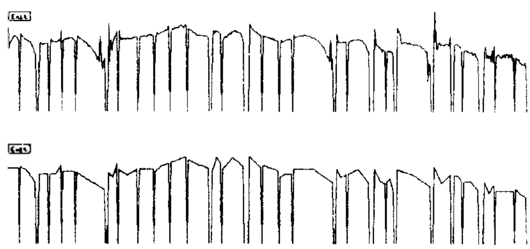

Figure 5: (a) Original intonation (b) Reconstructed intonation, in test sentence.

5 Conclusion

In this paper, we addressed the features of the fundamental frequency (*F0*) contour of Korean read speech, and proposed a new stylization method to characterize the *F0* pattern of segments. It was shown that our algorithm is robust to noisy pitch contour and makes reliable results. In addition, we could stylize pitch pattern of segment, syllable, and word with limited set which can reconstruct pitch contour with the error less than 10% of the actual *F0* contour for each sentence. In perception test, there was little difference between the synthesized speech with the original *F0* contour and the synthesized speech with the stylized *F0* contour.

References

[1] A.S.Madhukumar, S.Rajendran, and B.Yegnanarayana, "Intonation component of a text-to-speech system for Hindi," in *Computer Speech and Language*, Academic Press, pp.283-301, Jul. 1993

[2] G.Olaszy, G.Gordos, and G.Nemech, "The multivox multilingual text-to-speech converter," in *Talking Machines:Theories, Models, and Designs*, North-Holland, pp.385-411, 1992

[3] F.Emerard, L.Mortamet, and A.Cosannet, "Prosodic processing in a text-to-speech synthesis system using a database and learning procedures," in *Talking Machines:Theories, Models, and Designs*, North-Holland, pp.225-254, 1992

[4] J.Pierrehumbert "Synthesizing intonation," in *J. Acoust. Soc. Am.*, vol.70, no.4, pp.985-995, 1981

[5] S.H.Kim and J.C.Lee "Korean Text-to-Speech System Using TD-PSOLA," (to be published in SST-94)

[6] S.H.Kim and M.J.Zhi, et.al. "Application of TD-PSOLA Technique to Korean T-t-S Conversion," in *Proc. IWSP 93*, pp.83-88, Nov. 1993