A computational algorithm for F0 contour generation in Korean developed with prosodically labeled databases using K-ToBI system

[1]*Yong-Ju Lee*, [2]*Sook-hyang Lee*, [1]*Jong-Jin Kim*, [2]*Hyun-Ju Ko*, [1]*Young-Il Kim*

[3]*Sang-Hun Kim*, [3]*Jung-Cheol Lee*

[1]Dept. of Computer Eng., [2]Dept. of English Language and Literature
Wonkwang Univ. 344-2 Shinyong-Dong, Iksan , Chonbuk 570-749 KOREA
E-mail address : yjlee@wonnms.wonkwang.ac.kr, Fax number : (+82) 653-856-8009
[3]Spoken Language Processing Section
Electronics and Telecommunication Research Institute
P.O.BOX 106 Yu-Seong Post Office, Taejeon, KOREA

ABSTRACT

This study describes an algorithm for the F0 contour generation system for Korean sentences and its evaluation results. 400 K-ToBI labeled utterances were used which were read by one male and one female announcers. F0 contour generation system uses two classification trees for prediction of K-ToBI labels for input text and 11 regression trees for prediction of F0 values for the labels. Evaluation results of the system showed 77.2% prediction accuracy for prediction of IP boundaries and 72.0% prediction accuracy for AP boundaries. Information of voicing and duration of the segments was not changed for F0 contour generation and its evaluation. Evaluation results showed 23.5Hz RMS error and 0.55 correlation coefficient in F0 generation experiment using labelling information from the original speech data.

1. INTRODUCTION

Speech signal contains not only text-convertible information but also a variety of information such as speaker's emotion, intention or demand upon listeners. Many linguists/phoneticians analyze the speech signal into segmental and suprasegmental properties. Speech synthesis system, as an agent of speech communication tools for human beings, should be able to not only automatically generate segmental features but also mock as much prosodic features as possible. The ultimate goal of this study is to develop a synthesis system to produce Korean natural speech. In order for this, an algorithm is needed for both modeling segmental and prosodic features and adopting an optimal synthesis unit from the large size of speech database.

In order to adopt an optimal synthesis unit for input text from the speech database, target value generation module should be developed and it determines the quality of the synthesized speech.

This work was supported by Electronics and Telecommunication Research Institute.

Target value generation module consists of accurate prediction models of duration, F0 contours and energy contours. In order to develop target value generation system for F0 contours, this study made an attempt to develop F0 contour generation system using K-ToBI labeled speech database, and also evaluated its quality. It will be used as a target F0 generation module for deriving synthesis unit for natural speech and also used as a F0 correction module for the synthesized speech.

2. F0 CONTOUR PREDICTION MODEL

CART decision trees were designed for predicting K-ToBI labels for input text and F0 values for them based on the assumptions on the Korean prosodic structure discussed in [1][2][3] and K-ToBI labelling system presented in [3].

Korean hierarchical prosodic structure labeled using K-ToBI symbols is given in Figure 1.

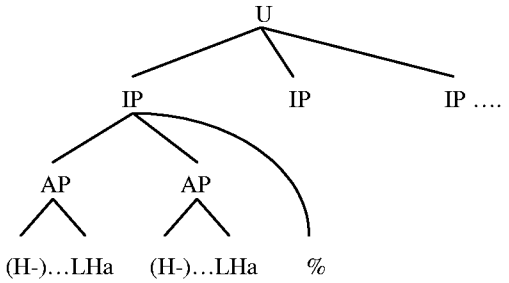

Figure 1: Korean hierarchical prosodic structure[3]

An Utterance(U) consists of one or more than one Intonational phrase(henceforth IP) and one IP consists of one or more than one Accentual phrases(henceforth AP). F0 range downstep among APs occurs within an IP and it is reset with a new IP. IP is demarcated by a boundary tone(e.g., H%) and AP by a rising tone(LHa). Description of a typical contour of AP in Korean and its tonal events is given in Figure 2. Their phonetic characteristics are discussed in [3] and [4]. Since occurrence of

'H-' depends on the number of syllables within AP, it could be realized in a different pattern as shown in Figure 3.

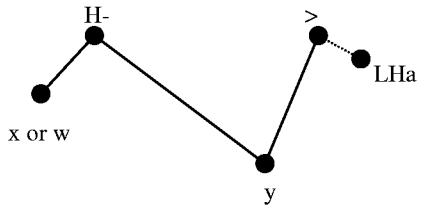

Figure 2: Typical pattern of Korean AP contour.

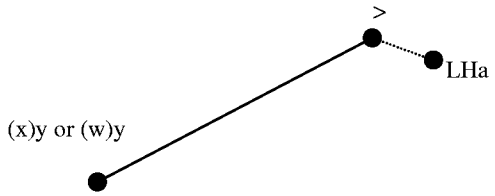

Figure 3: One of variant patterns of Korean AP contour.

2.2 Speech Data

400 K-ToBI labeled sentences were used out of 2,000 sentences read by one male and one female amateur announcers.

The sentences were adopted from a variety of sources such as radio news and also various with respect to their length and syntactic complexity.

2.3 Modeling procedure for F0 contour prediction

Procedure for F0 contour generation for the input text is described below.

1. Prediction of IP boundary location
2. Prediction of AP boundary location using information on IP boundary location
3. Prediction of syllable location for 'x/w', 'H-' ,'y', and '>'
4. Prediction of location for the predicted K-ToBI label within a syllable
5. Prediction of F0 value for each label
6. F0 contour generation using Spline algorithm

F0 generation algorithm consists of two models: prosodic label prediction model and F0 value prediction model.

2.4 Prediction Model for K-ToBI labels

2.4.1 Prediction of Intonational phrase boundary location

Through CART training using 10 features given in Table 1, Classification tree was generated for predicting IP boundary location.

Type	Description
Categorical feature	Preferred cluster of parts-of-speech for IP boundary word
	Previous word's part-of-speech
	Candidate word's part-of-speech
	Next word's part-of speech
Continuous feature	Number of words in Sentence
	Number of syllables in Sentence
	Number of words since the beginning of Sentence
	Number of syllables since the beginning of Sentence
	Number of words to the end of Sentence
	Number of syllables to the end of Sentence

Table 1: 10 features for predicting IP boundary location.

Boundary tone label was assigned to the last syllable of the last word of the given IP.

As a default type of the boundary tone, HL% was chosen.

HL% accounts for 48.1% of the boundary tone types from model training data, therefore decision tree grown by CART training could not predict other types except HL% type.

It might be because many of the sentences were selected from radio news and sentence type of all sentences was declarative.

2.4.2 Prediction of Accentual phrase boundary location

Classification tree for predicting AP boundary was grown through CART training using 13 features given in Table 2.

Type	Description
Categorical feature	Preferred cluster of parts-of-speech for AP boundary word
	Previous word's of part-of-speech
	Candidate word's part-of-speech
	Next word's part-of-speech
Continuous feature	Number of previous APs
	Number of words in Sentence
	Number of syllables in Sentence
	Number of words since the boundary of immediately preceding AP
	Number of syllables since the boundary of immediately preceding AP
	Number of words since the beginning of Sentence
	Number of syllables since the beginning of Sentence
	Number of words to the end of Sentence
	Number of syllables to the end of Sentence

Table 2: 13 features for predicting AP boundary location.

As a default type of AP boundary tone, 'LH' was chosen.

2.4.3 Prediction of position for labels, 'x/w', 'H-', 'y', '>' in an Accentual phrase

Syllables for 'x/w', 'H-' , 'y', and '>' were chosen based on the results of the statistical analysis on these tonal events done in [4]. Occurrence of 'H-' depends on the number of syllables of a given AP. 96.9% of APs with more than 3 syllables were realized with 'H-' while 86.2% of APs with less than 4 syllables did not show H- at the beginning. Thus, 'H-' was assigned only to AP with more than 3 syllables. It was assigned to the first syllable or the second syllable of a given AP according to the following rules.

```
if   [1st syllable is a super-heavy or heavy syllable] then
       assign 'H-' to 1st syllable
else if [1st syllable is a light syllable] then
       assign 'H-' to 2nd syllable
else if [1st syllable is a heavy-syllable and 2nd-syllable is a
         super-heavy syllable] then
       assign 'H-' to 2nd syllable
endif
```

When AP has 'H-', 'y' was assigned to the penultimate syllable while it was assigned to the antepenultimate syllable when AP did not have 'H-'.

'>' was assigned to the last syllable of AP.

2.4.4 Location of tonal events within a syllable

Once syllable for each K-ToBI label is determined, its exact position within a voiced period in a given syllable is determined. Position of each tonal event within a syllable was derived from the following formula.

$$R = \frac{D_2}{D_1} \leq 1.0 \qquad \text{(Eq. 1.)}$$

D_1 is duration of voicing period of a given syllable for each tonal event and D_2 is duration from the beginning of the syllable to the position of each tonal event within voiced period.

K-ToBI label	Location of tonal events within a syllable	
	Mean of R.	SD. of R.
X	.441	.292
W	.478	.319
H-	.535	.212
Y	.585	.343
>	.510	.240
LHa	.715	.444
(X)H-	.531	.188
(W)H-	.520	.221
(X)Y	.351	.263
(W)Y	.593	.260
HL%	.752	.330

Table 3: Location of tonal events within voiced period of a given syllable.

2.5 Model predicting F0 value for K-ToBI labels

Once K-ToBI labels are assigned to the appropriate position within a syllable, F0 value for each label is predicted. Model predicting F0 value for each prosodic label is composed of 11 regression trees. Features predicting F0 value are given below.

- Number of preceding and following IPs within Sentence
- Number of preceding APs within a given IP
- Number of syllables within AP

3. EVALUATION

Two experiments were conducted for evaluation of F0 contour generation system. In experiment I, generated F0 contours from prosodic label prediction model and F0 value prediction model were compared to those of original data. In Experiment II, in order to evaluate F0 contour generation module itself, F0 contours generated from F0 prediction module using label information of original data were compared to those of original data. There was no change in duration and energy contours in the synthesized data from original data.

3.1 Experiment I

In Experiment I, F0 contours generated from label prediction model and F0 prediction model were compared to those of original data. First, comparison of predicted IP boundary locations to those of original data was made. Results are given in Table 4.

	No IP Boundary	IP Boundary
No IP Boundary	85.9%	14.1%
IP Boundary	22.8%	77.2%

Table 4: Results of comparison of predicted IP boundary locations to those of original data

Comparison of predicted AP boundary locations to those of original data was also made. Results are given in Table 5.

	No AP boundary	AP boundary
No AP boundary	29.6%	70.4%
AP boundary	28.0%	72.0%

Table 5: Results of comparison of predicted AP boundary locations to those of original data.

3.2 Experiment II

In Experiment II, F0 contour generation module itself was evaluated by comparing F0 contours generated from F0 prediction module using label information of original data to those of original data.

RMS error and correlation coefficient of F0 value for each label from original data are given in Table 6.

K-ToBI label	F0 value of original data(Hz)	Predicted F0 value(Hz)	RMS Error	Corr.
X	133(29)	130(12)	29.8	.392
W	130(22)	127(9)	26.8	.337
H-	154(21)	152(12)	24.9	.546
Y	122(22)	121(13)	25.2	.605
>	149(23)	147(15)	24.7	.641
LHa	148(23)	144(13)	25.7	.522
(x)H-	168(21)	173(9)	26.2	.337
(w)H-	162(26)	157(15)	31.0	.449
(x)y	127(23)	129(10)	24.0	.446
(w)y	122(18)	121(10)	22.9	.499
HL%	91(28)	90(0)	23.1	.

Table 6: RMS error and correlation coefficient of F0 value for each label from original data (parenthesis indicates standard deviation).

RMS error and correlation coefficient of synthesized F0 contours from original data are shown in Table 7.

RMS Error	Correlation
23.56	.55

Table 7: RMS error and correlation coefficient of synthesized F0 contours from original data.

An example of F0 contours generated from Experiment I and II and those of original data are illustrated in Figure. 4.

4. CONCLUSION

This study made an attempt to develop a F0 contour generation algorithm using K-ToBI labeled speech data with an ultimate goal to produce natural synthesized speech in Korean. Its evaluation results were also reported. 400 sentences produced by one male and one female speakers were used for model training and system evaluation. Through CART training, classification tree predicting F0 value for each prosodic label and 11 regression trees were grown predicting F0 value for each label. Evaluation of the system was also done. It has several problems. Lack of stability of the model is one of them. It could be fixed by using much bigger size of data for model training.

Inaccuracy in prediction of IP boundary location is also a problem of the system. It might be fixed by adding pause information predicted by pause prediction model.

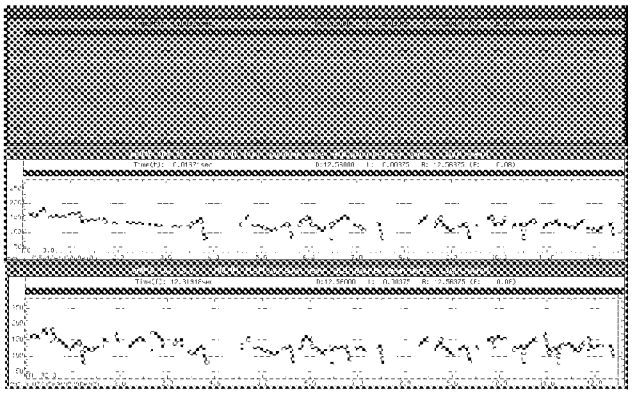

Figure 4: (top) F0 contours of original data; (middle) F0 contours generated from prosodic label prediction model and F0 value prediction model; (bottom) F0 contours generated from F0 value prediction model using prosodic label information of original data.

5. REFERENACE

1. Mary E. Beckman and Sun-Ah Jun, *K-ToBI(Korean ToBI) Labeling Conventions version 2.1,* Revised Nov. 1996.
2. Sun-Ah.Jun, *The Phonetics and Phonology of Korean prosody*: Ph.D. Dissertation, The Ohio State Univ., 1993.
3. Yong-Ju Lee, Sook-hyang Lee, Sang-Hun Kim, and Jong-Jin Kim, *A Study on developing the computational algorithm for generation of F0 contours using K-ToBI labelling system.*(in Korean), Electronics and Telecommunication Research Institute final Report, Wonkwang University, 1997.
4. Jong-Jin Kim, Sook-hyang Lee, Hyun-Ju Ko, Yong-Ju Lee, Sang-Hun Kim, and Jung-Cheol Lee, "An Analysis of some prosodic aspects of Korean utterances using K-ToBI labelling system," Proc. ICSP, Vol. 1, pp.87-91, 1997.
5. Kenneth N. Ross, *Modeling of Intonation for Speech Synthesis*: Ph.D. Dissertation, Boston Univ., 1995.
6. Alan W. Black, Andrew J. Hunt, "Generating F0 Contours from ToBI labels using linear regression," Proc. ICSLP , Vol. 3, pp. 1385-1388, 1996.