# INTONATION PROCESSING FOR TTS USING STYLIZATION AND NEURAL NETWORK LEARING METHOD

*Jung-Chul Lee, Youngjik Lee, Sang-Hun Kim, and Minsoo Hahn*

Electronics and Telecommunications Research Institute,
Yusong P.O. Box 106, Taejon, 305-350, Korea
E-mail:jclee@zenith.etri.re.kr

## ABSTRACT

In this paper, we propose a new model for synthesizing fundamental frequency (F0) contours using a stylization and a neural network learning method. The F0 contour is described as the superposition of 4 layered features; global tune, word pitch bias, lexical tone, and the syllabic pitch pattern. We firstly stylize the F0 contour of speech material, and analyze stylized data by statistical approach according to grammatical attributes. We then construct a melodic table, and train lexical tone with a neural network. Finally we develop the intonation generation rules for TTS conversion. This model produces a good neutral declarative intonation, and there is little difference between synthesized speech with original F0 contour and that with the rule generated contour when tested with our TD-PSOLA synthesizer[6][7].

## 1. INTRODUCTION

Intonation plays an important role in the intelligibility and naturalness of speech. Many researches have devoted their efforts to establish a formal representation of intonational function and form, or to its realization [1]-[3]. These researches treated the F0 contour as the phonetic accompaniment of certain types of syntatic units or constituent boundaries, and affecting functions to intonation. But these require a sophisticated parser, and the results are still unsatisfactory [4]. Recently, Emerard et al. proposed an intonation processing method in a TTS system based on the stylization method to yield natural synthetic speech [5]. The purpose of stylization is reduction of redundancy and easiness of manipulation for pitch patterns. He adopted contextual parsing rules to determine prosodic boundaries, and used a melodic table to provide F0 patterns for words. But his algorithm realized boundary effects only and did not consider the interaction between words in contextual effects.

To implement the interaction between words in contextual effects, we try to extend stylization to the non-uniform grammatical unit which can represent word, sequence of words, phrase, or sentence. For the effective analysis/synthesis of F0 contour, we make a new intonation processing model which assumes the F0 contour as the superposition of 4 layered features; 1) global tune which assigns potential pitch mean value to each word, 2) word pitch bias strongly related to the grammatical context, (word pitch bias means the F0 difference between the real mean F0 value and the predicted value in the declination line for each word) 3) lexical tone which means the mean F0 values of syllables in a word subtracted from by word mean F0 value, 4) the syllabic pitch pattern as a microprosodic component.

The main difference of our algorithm from other approaches is that the layer 1 and layer 2 assign one pitch value to each word, and the layer 3 assigns one pitch value to each syllable. The advantage of this idea is that we can easily implement nonlinear characteristic of F0 baseline with a statistical mapping table, i.e., the melodic table, between the grammatical context and word pitch bias. Also this idea enables the linear analysis/synthesis of the F0 contour. For the analysis of grammatical context, we define 60 grammatical attributes which can be deduced by particles, suffix inflections, adverbs, and conjunctions.

As a global tune, we use a linear declination line, y=at+b, where t is normalized by the number of words in a sentence. We estimate a and b by applying the LS method for our 24 minutes' speech material of 156 sentences. With the word pitch biases and grammatical attributes for words, we construct our melodic table through statistical approach.

Korean is neither a stressed nor a tonal language because the pitch accent in a word is not culminative and different tone does not alter the meaning of a word. But if the word pitch pattern(lexical tone) is not proper, it sounds like a dialect or a foreign accent. We observed that lexical tone has many regularities in real speech and a 2-layer neural network can predict lexical tone from the phonetic and phonological information. Therefore, we adopted a neural network to train lexical tone. And finally we constructed the intonation generation rules for TTS conversion.

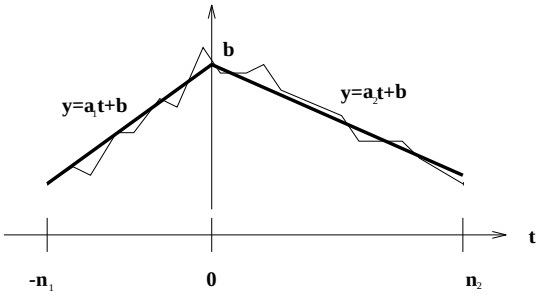


Figure 1: Two straight line approximation

## 2. STYLIZATION OF PITCH PATTERN

Generally, the estimate of pitch for speech signal has fluctuations on pitch contour due to the estimation error or unstable vibration of the vocal folds. To minimize these defects in the stylization of the syllabic F0 pattern, we employed the least square error minimization method to estimate the F0 values at three positions. Firstly, the syllabic pitch contour was piecewise-linearly approximated by two straight lines, as shown in Figure 1. 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, we can calculate new pitch values and the mean F0 for the segment. Also, the position of peak or valley can be estimated if we find the minimum of LS errors by varying $n_1$ and $n_2$. So this method can automatically detect pitch movement timing in a segment. Then the syllabic F0 pattern is stylized by using four features; the pitch movement timing normalized by syllable duration, three point pitch values which are chosen at $-n_1, 0, n_2$. And we subtract mean F0 value of each syllable from all pitch informations. With the stylized data in our database, we find the representative styles for all Korean syllables. Korean syllable consists of CVC, where initial C and final C can be deleted. In Korean, there exist 18 initial C, 21 V, and 7 final C.

We define lexical tone as the F0 values of syllables in a word subtracted from the word mean F0 value. We use a 2-layer neural network, as shown in Figure 2, to produce lexical tone directly from encoded phonological representation. The phonological representation are categorized by 25 factors according to manner, position and intensity of articulation; 15 for consonant and 10 factors for vowel, respectively.

Finally, the global tune is estimated from the mean F0 values of words in a sentence by using the LS method.

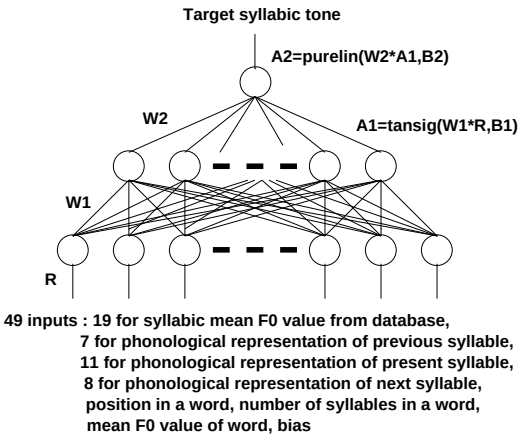


Figure 2: Structure of the neural network for lexical tone prediction

## 3. INTONATION GENERATION MODEL

In the development of an intonation generation model, we assume that the intonation in a sentence can be characterized by the following process. 1) The primary factor which assigns the potential pitch mean value to each word, neither segments nor syllable, is the position in a sentence, and the value can be determined by the global level tune, y=at+b. 2) This declination line assigns the potential pitch mean value to each word according to the primary factor, i.e., the position in a sentence. 3) The secondary factor which determines the bias and the pitch pattern for each word is grammatical attributes. 4) The number of neighboring words which affects the bias and pitch pattern is $1 \sim 5$. In other words, the primary factor for assigning a pitch value to each word is the position of that word in a sentence, and the secondary factor is the grammatical attribute for each word. And in addition, the secondary factor produces an absolute, not relative, pitch value. Most of intonation generation models usually use parsers to analyze the syntactic structure of sentence, and the parsers become more sophisticated to guarantee the good results. But our algorithm does not depend on a complex parser. We, instead, try to build a direct mapping table between the secondary factor and the word pitch bias. If there is a well-defined set of grammatical attributes and these are carefully derived, the mapping table can be successfully built through the statistical approach. The mapping table construction procedures are as follows.

$\overline{F0}_{ij}$ and the pitch pattern, $P_{ij}$, of each word are obtained through the stylization process, where i is for the sentence number and j is for the position of the word. Using $\overline{F0}_{ij}$, the representative baseline can be found by applying the least squared error method which is normalized with respect to the number of words in a sentence. Then we calculate $d_{ij}$, the difference between $\overline{F0}_{ij}$ and the baseline, and tag one grammatical attribute $g_{ij}$ to each word among the set

$\{G_k|k = 1, K\}$. Let $\bar{\Delta}n$ be $\bar{\Delta}_{G_{n_1}}\bar{\Delta}_{G_{n_2}}\cdots\bar{\Delta}_{G_{n_n}}$, where $\bar{\Delta}_{G_{n_k}}$ is equal to $\frac{1}{N_n}\sum_{g_{ij}=G_{n_k}} d_{ij}$ . We first find $\bar{\Delta}n$ for $n = 1, \cdots, 5$, and remove redundancies if $\bar{\Delta}n$ can be explained by two $\bar{\Delta}(n-1)$. The deletion criterion is

$$E = \sum_{i=1}^{n-1}(\bar{\Delta}_{G_{n_i}} - \bar{\Delta}_{G_{(n-1)_i}})^2 + \sum_{i=2}^{n}(\bar{\Delta}_{G_{n_i}} - \bar{\Delta}_{G^c_{(n-1)_i}})^2 < threshold. \quad (2)$$

As a result, the melodic table is constructed with the pitch bias sequences according to grammatical attribute sequences.

The intonation processing module in our TTS system calculates the primary pitch value according to the word position in a sentence, and searches for the longest matching entry by utilizing the information from the grammatical attribute sequence. If the best-matching entry is selected, the pitch bias for that word is extracted from the table. With the phonological representation and the sum of primary and secondary pitch values, we can calculate the mean F0 value for each syllable in a word. In the realization of syllabic pitch pattern, we use reference patterns determined by syllable types. We finally calculate the F0 contour of a sentence using 2nd order parabolic interpolation.

## 4. SIMULATION RESULTS

We have performed four experiments. The speech material for our present study consists of recordings produced by reading individual sentences of the written text separately. The written text is composed of 156 sentences having 2186 words and 6632 syllables, and read by an experienced female announcer. The recorded material is digitized at 16 KHz with 16 bit resolution. The fundamental frequencies are extracted by using the ESPS S/W, while the markers for segments are manually labeled.

Firstly, we test the fitness of a linear declination line as a global tune. Figure 3 shows the F0 contour patterns of 156 sentences and a linear line y=at+b, where t is normalized by the number of words in a sentence, is proved to be good enough to represent the global tune.

Secondly, we find the pitch bias for each word in a sentence using the melodic table according to the grammatical attribute sequences, and calculate the prediction error as shown in Figure 4. If we assign a F0 value of global tune to each word in the layer 1 processing, the standard deviation (SD) is 18.7 Hz. After the layer 2 processing, we can reduce SD to 1.6 Hz. This implies that our algorithm is very simple but effective to accomodate the interaction between words in contextual effects.

Third one is the training of lexical tone with a neural network. Figure 5 shows prediction error for lexical tone. We have experimented with various learning rates, ranged from

| number of syllables in a word | number of words | SD of prediction error after layer 2 [Hz] | SD of prediction error after layer 3 [Hz] |
|---|---|---|---|
| 2 | 573 | 9.2 | 1.1 |
| 3 | 829 | 11.8 | 2.8 |
| 4 | 401 | 13.2 | 1.8 |
| 5 | 188 | 14.6 | 1.1 |
| 6 | 48 | 14.9 | 0.0 |
| 7 | 16 | 14.3 | 0.0 |
| 8 | 2 | 12.6 | 0.0 |
| 9 | 3 | 18.1 | 0.0 |
| 12 | 1 | 17.8 | 0.0 |
| total | 2061 | 12.4 | 2.0 |

Table 1: Prediction error from the neural network

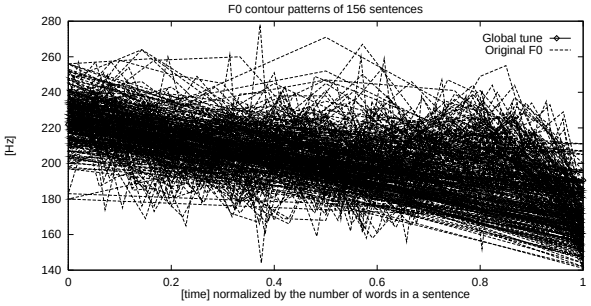


Figure 3: F0 contour patterns of 156 sentences and Global tune

10 to 0.00001 with initial weights of very small random values. The results show that if the learning rate is not sufficiently small, learning process fails. So we use 0.0001 for both layers. Table 1 shows the prediction error. These results came from the independent neural networks with respect to the number of syllables in a word. And we admit that we need more data of words having more than 6 syllables to generalize our results. Still our results can be a strong proof that the neural network can be very effective to predict the lexical tone, and to improve the naturalness of the synthesized speech.

Figure 6 shows both the original and the rule generated pitch contour. The mean absolute deviation between the original contour and the rule generated F0 contour was 11.4Hz. In other words, the rule generated F0 pattern traces the original F0 contour fairly well. In the perception test, little difference was detected between the synthesized speech with the original F0 contour and that with the synthesized one.

## 5. CONCLUSION

In this paper, we proposed a new model for synthesizing F0 contours using a stylization and a neural network learning

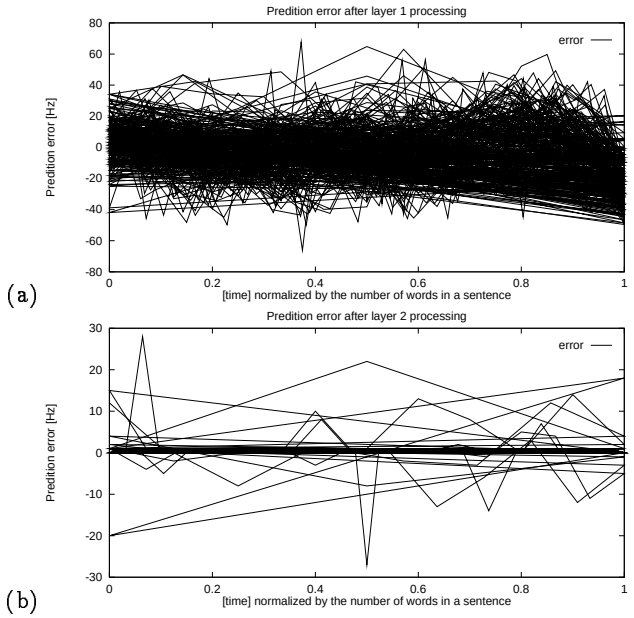


Figure 4: Prediction errors after (a) layer 1, (b) layer 2.

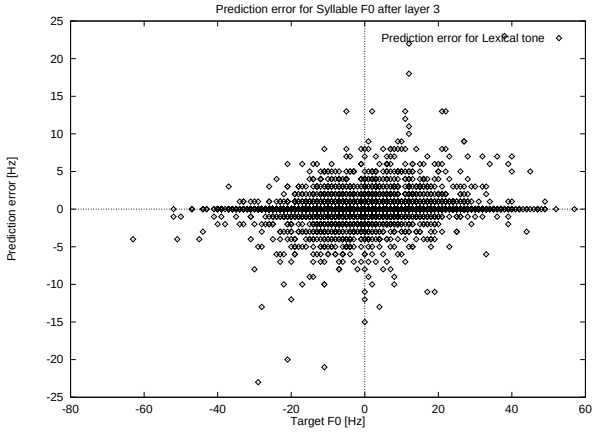


Figure 5: Prediction error for lexical tone

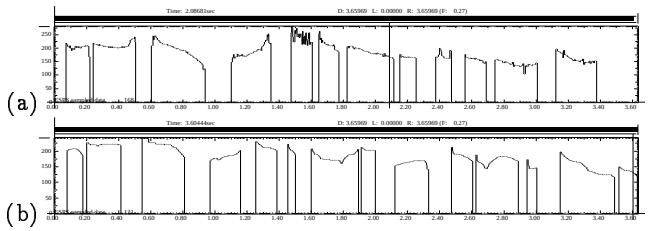


Figure 6: Comparison of F0 contour for test sentence (a) Original intonation (b) Rule generated intonation. Time alignment is not carried out in this experiment.

method. The simulation results of each layer show our algorithm is very effective to analysis/synthesis the F0 contour of a sentence, and generates a fairly good declarative intonation. Our future works include the improvement of the prediction accuracy for the lexical tone with a large speech database. Also we will try to do some pretests in search of the methods of extending this model to accomodate dialogue-type speech.

# References


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

[2] D.Hirst "Structures and Categories in Prosodic Representation," in *Prosody: Models and Measurements*, Springer-Verlag, pp.93-109, 1983

[3] H.Fujisaki and K.Hirose "Analysis of voice fundamental frequency coutours for declarative sentences of Japanese," in *J. Acoust. Soc. Jpn.*, 5(4), pp.233-242, 1984

[4] D.Hirst "Prediction of prosody: An overview," in *Talking Machines:Theories, Models, and Designs*, North-Holland, pp.199-204, 1992

[5] F.Emerard, L.Mortamet, and A.Cozannet, "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

[6] J.C.Lee, and S.H.Kim, and Minsoo Hahn "Intonation Processing for Korean TTS Conversion Using Stylization Method," in *Proc. ICSPAT95*, pp.1943-1946, 1995

[7] S.H.Kim and J.C.Lee "Korean Text-to-Speech System Using TD-PSOLA," in *Proc. SST94*, pp.587-592, 1994