An Event-Driven f_0 Weighting for Prosody Control in a Large Corpus-Based TTS System

Heo-Jin Byeon and Yung-Hwan Oh, *Member, IEEE*

***Abstract*—This letter presents event-driven weighting methods concentrating on f_0 for prosody control in a large corpus-based text-to-speech (TTS) system. We determine the f_0 weighting factor for a given target using its linguistic features by automatically using classification and regression trees (CART). The target predicted as perceptually important is weighted more than others. This results in more natural synthetic speech from a prosodic viewpoint.**

***Index Terms*—CART, event-driven f_0 weighting, large corpus-based TTS system.**

I. Introduction

RECENTLY, large corpus-based concatenative synthesis has been the most popular approach for constructing TTS systems. With this method, it should be possible to synthesize more natural sounding speech than can be produced with a small set of controlled units. The selection of appropriate units from a given corpus for synthesis is based on two cost functions, the concatenation cost and the target cost. The concatenation cost is an estimate of the quality of a join between consecutive units, and the target cost is an estimate of the difference between a unit in the given corpus and a desired target [1].

In particular, the target cost plays an important role in prosody control. Prosody aids the listener in interpreting an utterance: therefore, it is considered as an important factor for natural sounding synthetic speech. To reflect prosody, the differences of acoustic parameters such as f_0 and duration have been used as a part of the target cost. From this point forward in this letter we refer to these differences in parameters as prosody cost. The use of perceptually motivated prosodic categories, ToBI Lite, as the prosody cost has also been explored to improve the naturalness of synthetic speech [2]. Generally, desired targets are predicted based on the linguistic information of a given sentence. The targets can be classified into perceptually more important ones or less important ones in synthetic speech. Consequently, it would be more effective in terms of prosody control to weight the prosody cost according to the degree of the perceptual role of targets in a given sentence.

We describe five methods for weighting the prosody cost. The methods reflect the perceptual importance of targets in a given sentence using CART [3], concentrating on f_0.

Manuscript received January 6, 2003; revised May 17, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

The authors are with the Spoken Language Laboratory, Computer Science Division, Korea Advanced Institute of Science and Technology, Daejon 305-701, South Korea (e-mail: bhjin@bulsai.kaist.ac.kr).

Digital Object Identifier 10.1109/LSP.2003.821684

II. Event-Driven f_0 Weighting Methods

According to the Tilt model [4], which is a phonetic model of intonation proposed by Paul Taylor, the f_0 contour of a given utterance can be represented as a sequence of intonational events affecting the perception of intonation such as pitch accents and boundary tones. Events occur as instants with nothing between them, as opposed to segmental based phenomena where units occur in a contiguous sequence. Thus we need only the f_0 contour of events and the other part of an f_0 contour can be interpolated using the boundary f_0 values between consecutive events. Using this property, the f_0 difference between the target predicted as an event and a candidate unit is weighted more than that of the other cases. Consequently, the f_0 values of perceptually important targets are emphasized and those of the others have interpolation effects by the f_0 difference between consecutive units as a part of concatenation cost.

We use a decision tree to predict whether a target is an event and a regression tree to predict the f_0 value of a target. Given a training set $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, a decision tree and a regression tree are made by successively dividing the regions of feature vector $\mathbf{x}$, the linguistic features of a target in our study, to minimize prediction error. When an unknown feature vector is submitted, the tree estimates its response by outputting the value of the terminal node that it lands in [3].

We define the f_0 weighting factor as a function of the linguistic feature vector $\mathbf{l}$ of a given target as follows:

$$W_{f_0}(\mathbf{l}) = w_{f_0} \cdot P_{f_0}(\mathbf{l}) \tag{1}$$

where w_{f_0} is the relative f_0 weighting factor against the target cost, which can be determined using the spectral differences between units and multiple linear regression [1], and $P_{f_0}(\mathbf{l})$ is the importance measure of a given target for the f_0 perception. Defining $P_{f_0}(\mathbf{l})$ is the key to achieving naturalness of synthetic speech. We propose five measures for $P_{f_0}(\mathbf{l})$.

At first, we use the result of the decision tree as a measure of $P_{f_0}(\mathbf{l})$. The output of the decision tree, $d_b(\mathbf{l})$, can be defined as

$$d_b(\mathbf{l}) = \begin{cases} 1, & \text{if } \mathbf{l} \text{ is classified as an event} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Using $d_b(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ results in reducing overall relative importance of f_0 in the target cost, because $d_b(\mathbf{l})$ is 0 in many cases. To solve this problem, we introduce a compensation factor, D_c, as

$$D_c = \frac{N_T}{N_e} \tag{3}$$

1070-9908/04$20.00 © 2004 IEEE

where N_e is the number of event instances and N_T is the number of samples in the decision tree. Then, we define $P_{\text{binary\_}f_0}(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ using (2) and (3) as follows:

$$P_{\text{binary\_}f_0}(\mathbf{l}) = D_c \cdot d_b(\mathbf{l}). \tag{4}$$

In this case, equal weights are assigned to the targets predicted as an event. However, different weights can be assigned even among the targets predicted as an event according to their importance. That is, the target predicted as a more obvious event is given more weights than a less obvious one.

We, therefore, redefine the response predicted by the decision tree given $\mathbf{l}$ as

$$d_c(\mathbf{l}) = \frac{E\left(t(\mathbf{l})\right)}{N\left(t(\mathbf{l})\right)} \tag{5}$$

where $t(\mathbf{l})$ is the terminal node of the decision tree that $\mathbf{l}$ lands in, $N(t)$ is the number of samples in node t and $E(t)$ is the number of event instances in node t. We then define $P_{\text{continuous\_}f_0}(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ using (3) and (5) as

$$P_{\text{continuous\_}f_0}(\mathbf{l}) = D_c \cdot d_c(\mathbf{l}). \tag{6}$$

The compensation factor D_c applied in (4) and (6) scales $P_{f_0}(\mathbf{l})$ to balance $W_{f_0}(\mathbf{l})$ against the other target costs. D_c makes $P_{f_0}(\mathbf{l})$ higher than 1 when the output of the tree is higher than the average event occurrence ratio and makes it lower than 1 otherwise. $d_b(\mathbf{l})$ can be considered as a special case of $d_c(\mathbf{l})$: $d_b(\mathbf{l})$ makes a hard decision while $d_c(\mathbf{l})$ makes a soft decision. Consequently, we incorporate the concept of scaling in the compensation factor, D_c, in (3), considering the average event occurrence ratio.

The above two weighting methods focus on events rather than their f_0 value. In our study, we predict the f_0 value of targets using a regression tree. The response of a regression tree, $d_r(\mathbf{l})$ is defined as the average of the responses of training feature vectors arriving at the terminal node t that $\mathbf{l}$ lands in, as follows:

$$d_r(\mathbf{l}) = \bar{y}\left(t(\mathbf{l})\right) = \frac{1}{N\left(t(\mathbf{l})\right)} \sum_{\mathbf{x}_n \in t(\mathbf{l})} y_n. \tag{7}$$

And the prospective prediction error, $e_r(\mathbf{l})$, is defined as follows:

$$e_r(\mathbf{l}) = \sqrt{\frac{1}{N\left(t(\mathbf{l})\right)} \sum_{\mathbf{x}_n \in t(\mathbf{l})} \left(y_n - d_r(\mathbf{l})\right)^2}. \tag{8}$$

We use $e_r(\mathbf{l})$ for weighting, because the target predicted with a small prospective prediction error is more reliable than that with a large error. Consequently, the reciprocal of $e_r(\mathbf{l})$ can be used as $P_{f_0}(\mathbf{l})$. In this case, we need a compensation factor to avoid changing the relative importance of f_0 in the target cost caused by applying the reciprocal of $e_r(\mathbf{l})$. The compensation factor R_c is defined as

$$R_c = \sqrt{\frac{1}{N_T} \sum_{n=1}^{N_T} \left(y_n - d_r(\mathbf{x}_n)\right)^2}. \tag{9}$$

Then, we define $P_{\text{regression\_}f_0}(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ using (8) and (9) as

$$P_{\text{regression\_}f_0}(\mathbf{l}) = \frac{R_c}{e_r(\mathbf{l})}. \tag{10}$$

The weighting factor determined using the regression tree can be combined with that using the decision tree. That is, the target regarded as more reliable and perceptually important is weighted more. We define $P_{\text{binary\_regression\_}f0}(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ using (4) and (10) as

$$P_{\text{binary\_regression\_}f0}(\mathbf{l}) = D_c \cdot R_c \cdot \frac{d_b(\mathbf{l})}{e_r(\mathbf{l})} \tag{11}$$

and define $P_{\text{continuous\_regression\_}f0}(\mathbf{l})$ as a measure of $P_{f_0}(\mathbf{l})$ using (6) and (10) as

$$P_{\text{continuous\_regression\_}f0}(\mathbf{l}) = D_c \cdot R_c \cdot \frac{d_c(\mathbf{l})}{e_r(\mathbf{l})}. \tag{12}$$

The feature vector used in the decision tree and that used in the regression tree are different from each other, but both represent the linguistic information of the target. Therefore, we used the same notation, $\mathbf{l}$, in both cases.

III. Experimental Results

We used a speech corpus comprising Korean utterances of a professional female speaker with a triphone as a basic unit for concatenation. The corpus comprised about 320 000 triphone instances. We used the Mel-frequency cepstral coefficient (MFCC), f_0 and energy distances between consecutive units as the concatenation cost, and f_0, duration and some phonological (the position of a unit in an utterance, etc.) distances between a candidate unit and the desired target as the target cost. Selected units were concatenated without signal processing.

The decision tree predicting an event occurrence and the regression tree predicting f_0 are implemented using Chou's algorithm [5] and the tenfold cross-validation method. In our study, we use the decision tree only to decide whether a given target is an event, regardless of its type. The proposed f_0 weighting methods concentrate on the obviousness and reliability of an event occurrence. Event types are necessary only when determining which f_0 pattern can be considered as an event in training data: an event is a pitch accent or a boundary tone as defined in [6]. To predict an event placement aligned with syllables, we used the following features (note that the initial alphabet of the feature denotes the type: D for categorical variable and C for real-valued one).

- Dnucleuspos, Dcodapos: The part of speech of the nucleus and coda in the syllable.
- Dsylltype: The configuration of the syllable. The feature takes one of four categories, N, ON, NC, or ONC (O: onset, N: nucleus, C: coda).
- Dsylllocphr, Dsyllloceoj: The locations of the syllable in the prosodic phrase and the *eojeol* (which is delimited by the spaces in a Korean sentence). The feature takes one of five categories, single, first, second, final, or rest.

- Dpunctype: The following punctuation mark of the eojeol that the syllable belongs to. The feature takes one of five categories, ",", ".", "?", "!", or "etc".
- Cphrlen, Ceojlen: The length of the prosodic phrase and eojeol in syllables.
- Cbsyllphr, Cesyllphr, Cbsylleoj, Cesylleoj: The syllable indexes from the beginning and the end of the prosodic phrase and the eojeol.
- Cbsyllphrr, Cbsylleojr: Cbsyllphr/Cphrlen, Cbsylleoj/Ceojlen.

We trained the tree on 500 utterances and tested it on 100 utterances. The training and test data include 9683 and 1609 syllables, respectively. The performance of the tree is shown in Table I.

We used a single f_0 value for a phone. But an event has one to three f_0 values because a Korean syllable has one to three phones. Therefore, an event can be regarded as being represented as a simplified f_0 contour. In our study, f_0 is represented on a semi-tone (ST) logarithmic scale. The 1-Hz frequency is taken as reference. Then the frequency in ST is obtained as $12 \cdot \log_2(f)$, with f expressed in Hertz. To predict the f_0 value for a given target, we used the following features.

- Dlph, Dcph, Drph: The preceding phone, the observed phone, and the following phone. These features represent a phone context.
- Dlpos, Dcpos, Drpos: The part of speech context corresponding to the phone context.
- Dphsylltype: The configuration of the syllable which the phone belongs to.
- Dphpuncloc: The following punctuation mark of the eojeol which the phone belongs to.
- Dphlocphr, Dphloceoj: The locations of the phone in the prosodic phrase and the eojeol. The feature takes one of four categories, single, initial, final or mid.
- Dphphrloc: The locations in a sentence of the prosodic phrase that the phone belongs to. The feature takes the same categories as Dphlocphr.
- Cphphrlen, Cpheojlen: The length of the prosodic phrase and eojeol in phones.
- Cbphphr, Cnphphr, Cbpheoj, Cnpheoj: The phone indexes from the beginning and the end of the prosodic phrase and the eojeol.
- Cbphphrr, Cbpheojr: Cbphphr/Cphrlen, Cbpheoj/Ceojlen.

We trained and tested the tree with the same data used for the decision tree. The performance of the tree is shown in Table II.

A mean opinion scores (MOS) test was conducted to determine whether the proposed weighting methods improve perceived synthetic speech quality. Six experimental f_0 weighting conditions were evaluated in the MOS test.

- *BL*: w_{f_0} in (1) was used as the f_0 weighting factor, which is trained to optimize the mapping from the feature vector differences between units to the MFCC distances between units in the corpus.
- *BE*: $W_{f_0}(\mathbf{1})$ defined in (1) with $P_{\text{binary\_}f_0}(\mathbf{1})$ as a measure of $P_{f_0}(\mathbf{1})$ was used as the f_0 weighting factor.
- *CE*: $W_{f_0}(\mathbf{1})$ with $P_{\text{continuous\_}f_0}(\mathbf{1})$ was used.
- *RF*: $W_{f_0}(\mathbf{1})$ with $P_{\text{regression\_}f_0}(\mathbf{1})$ was used.

TABLE I
PERFORMANCE OF THE DECISION TREE FOR EVENT PREDICTION

	Train (%)	Test (%)
Event	89.03 (3537/3973)	84.78 (507/598)
None	85.45 (4879/5710)	81.01 (819/1011)
Total	86.92 (8416/9683)	82.41 (1326/1609)

TABLE II
PERFORMANCE OF THE REGRESSION TREE FOR f_0 PREDICTION ($R(T)$: THE MEAN SQUARED ERROR OF TREE T)

Train	Test
$\sqrt{R(T)} = 1.39$	$\sqrt{R^{ts}(T)} = 1.88$

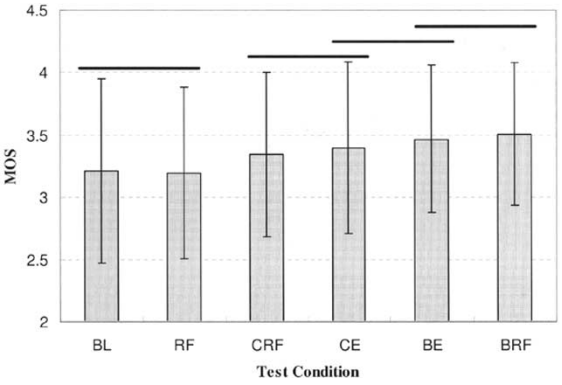

Fig. 1. Results of the MOS test.

- *BRF*: $W_{f_0}(\mathbf{1})$ with $P_{\text{binary\_regression\_}f_0}(\mathbf{1})$ was used.
- *CRF*: $W_{f_0}(\mathbf{1})$ with $P_{\text{continuous\_regression\_}f_0}(\mathbf{1})$ was used.

Twenty listeners participated in an MOS test consisting 30 test sentences, which means there were 180 synthetic utterances and 3600 observations. The listeners were adult Korean speakers with no known speech or hearing deficits. They were tested in four groups of 5 listeners. Each group was presented with a different random order of test utterances.

Fig. 1 shows the results of the MOS test. In the F test of ANOVA [7], we obtained $f = 22.26$ from our test data and $F_{0.05,5,3594} = 2.21$. These results indicate that there was a statistically significant difference among our test conditions at significance level $\alpha = 0.05$. In addition, Newman–Keuls tests [7] with $\alpha = 0.05$ were performed to compare the test conditions. The results are shown in Fig. 1: the horizontal lines above the bars indicate conditions whose ratings were not significantly different from each other. Unexpectedly, RF was rated lower than BL; however it was not significantly different. The other proposed methods were rated higher than BL with statistical significance. The results indicate that the event-driven f_0 weighting is effective. CRF was rated lower than CE, and BRF was rated higher than BE. These were not significantly different, respectively. However, BRF was rated higher than CE, and CRF was rated lower than BE with statistical significance while BE and CE were not significantly different from each other. These results indicate that the weighting factor combined with f_0 regression has positive effects in the binary event decision, but not in the continuous event decision. BE and BRF were rated higher

than the others. In addition, BE and BRF had smaller standard deviation than the others. That is, BE and BRF had fewer very low ratings than the others. This indicates that the simple binary event decision is better than the complicated decision.

IV. Conclusion

In this letter, we proposed five f_0 weighting methods for prosody control in a large corpus-based TTS system. The f_0 weighting factor was automatically determined by the linguistic features of a given target using CART. The effectiveness of the proposed event-driven f_0 weighting methods has been confirmed by experiments. In the future, we will apply an event-driven weighting method to other prosody elements such as duration and energy.

References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis using a large speech database," in *Proc. ICASSP*, 1996, pp. 1:373–1:376.

[2] C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis," in *Proc. ICSLP*, 2000, pp. 2:71–2:74.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. San Francisco, CA: Wadsworth, 1984.

[4] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *J. Acoust. Soc. Amer.*, vol. 107, pp. 1697–1714, 2000.

[5] P. A. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 340–354, Apr. 1991.

[6] S. H. Lee and Y. H. Oh, "Generating korean f_0 contour using cart," in *Proc. ICSP*, 1999, pp. 177–182.

[7] B. Winer, D. R. Brown, and K. M. Michels, *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1991.