USE OF PITCH PATTERN IMPROVEMENT IN THE CHATR SPEECH SYNTHESIS SYSTEM

Ken Fujisawa, Toshio Hirai, and Norio Higuchi*

e-mail: fujisawa@itl.atr.co.jp
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN

ABSTRACT

A corpus-based concatenative speech synthesis system using no signal processing can produce intelligible synthetic speech maintaining original voice characteristics, but it can sometimes be difficult to realize natural prosody. In such a concatenative system, it is very important to select appropriate waveform segments that are naturally close to the target prosody. This paper describes some approaches to unit selection for improving the prosody, especially intonation of such synthetic speech. If the unit selection measures for the fundamental frequency (F_0) are insufficient, the concatenative system may produce speech having a discontinuous F_0 pattern. Our proposed solution to this problem is to add extra measures for selecting units that form a smoother, more continuous F_0 contour. Through subjective experiments, we confirmed that each of these measures effectively improved intonation naturalness.

1. INTRODUCTION

The speech resequencing system CHATR [1][2] produces synthetic speech by concatenating phoneme-size waveform units from a natural speech database. Currently, no signal processing is done on the synthetic speech, so that it can preserve the voice characteristics of the original speaker. Prosodic features such as F_0 pattern and duration are used for unit selection by making a comparison with the target features predicted for an input utterance. However, if suitable units are not found, the intonation of the synthetic speech can be unnatural. This is due to two reasons: one is the inappropriate prediction of target F_0 pattern, and the other is inadequate unit selection. Since the former is a problem of F_0 prediction module, we consider the latter. Referring to the F_0 pattern, the current CHATR implementation stores only one value to represent the mean F_0 of each phoneme in the speech database, so it can be difficult to select appropriate phonemes that realize the exact target prosody. Applying signal processing such as PSOLA [3] is one solution to this problem, but this tends to cause distortion in the synthetic speech and reduces the voice characteristics of the original speaker. Consequently, instead of introducing signal processing, we propose the addition of an extra parameter to model the F_0 slope as well as some other measures that can select appropriate phonemes to realize the exact target F_0 pattern.

*Presently, Osaka Gas Information System Research Institute.

2. UNIT SELECTION

CHATR produces synthetic speech $u^n = (u_1, ..., u_n)$ from phonemes in the speech corpus by minimizing two distortion measures [1][2]. One is a target cost $C^t(t_i, u_i)$, and the other is a concatenation (concat) cost $C^c(u_{i-1}, u_i)$. The target cost $C^t(t_i, u_i)$ represents the distance between a target segment t_i and a candidate unit u_i in the speech corpus, *i.e.*, a weighted sum of the difference between the candidate unit features and the target segment features $C_j^t(t_i, u_i)$. The target cost $C^t(t_i, u_i)$ is shown as follows:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \qquad (1)$$

where p is the dimension of the feature vector. The feature vector consists of thirty prosodic and phonetic factors, including duration, power, and F_0 at the middle of each phoneme.

The concat cost $C^c(u_{i-1}, u_i)$ represents the distance between a selected unit and the adjoining unit previously selected and is defined as the sum of the difference between the two-unit feature $C_j^c(u_{i-1}, u_i)$ weighted by w_j^c. The concat cost $C^c(u_{i-1}, u_i)$ is shown as follows:

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \qquad (2)$$

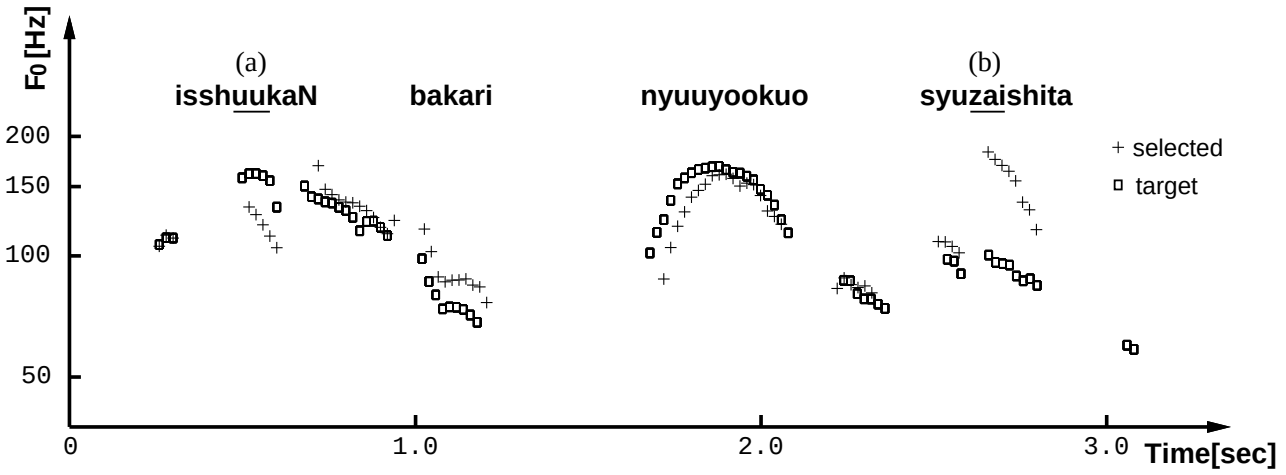

Figure 1: Example F_0 patterns of target and synthesized by CHATR.

Utterance: "isshuukaNbakari nyuuyookuo shuzaishita" (I collected news materials in New York for a week.) The '+' marks show the F_0 pattern of synthetic speech from CHATR. The boxes show target F_0, which in this case is derived from the original speech. The annotation (a), (b) is explained in the body of the text.

where q is the dimension of the feature vector. At this point, concat subcost consists of the following:

- cepstrum distance,
- difference of log power, and
- difference of F_0.

If the two units are adjacent phonemes of a speech file in the speech database, the concat cost is zero.

The total cost of n units is the summation of the target cost and the concat cost. The best sequence of units $\overline{u}^n$ is determined by minimizing the total cost.

$$\overline{u}^n = \underset{u_1,...,u_n}{\operatorname{argmin}} C(t^n, u^n) \qquad (3)$$

where

$$C(t^n, u^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i). \qquad (4)$$

3. ADDITIONAL SUBCOSTS FOR UNIT SELECTION

In addition to F_0 pattern, other features are used to characterize units. Hence, the selected units do not always have the ideal F_0 pattern for the target. Figure 1 illustrates the F_0 patterns of target and synthetic speech produced by CHATR. Here, the durations of each selected phoneme are aligned to the target duration for convenience. In this figure, the selected phonemes 'uu' (annotated as (a) in the figure) has a declining F_0 pattern ('+') though the target F_0 ('□') is flat, and the mean F_0 value of the selected unit is lower than the target. This causes an incorrect accent to be perceived in this region. The selected phonemes 'zai' (annotated as (b)) have higher F_0 values than the target. The shape of both F_0 patterns declines, but the mean F_0 of selected phonemes is so high compared with previous phonemes that it sounds overly stressed. As such, there should be distortion measures to evaluate differences, and the relative difference between adjacent units should be considered to realize appropriate intonation.

We added the following three cost functions to make the entire F_0 pattern faithful to the target F_0 pattern.

Slope cost denotes the difference in F_0 slope between candidate phoneme units and target segments. This is defined as a subcost of the target cost. The slope cost is considered only for vowels. To reduce any negative influence of F_0 extraction errors, the F_0 slope is calculated from linear regression after smoothing [4].

Threshold cost denotes the penalty cost applied when the difference in F_0 values between a candidate phoneme unit and target segment is larger than a threshold ($F_{0\,th}$). The penalty cost is larger than other costs, so selected phonemes should have F_0 values within the threshold from target F_0. This is also a target subcost and it is defined as follows:

$$C_j^t(t_i, u_i) = f(t_i - u_i) \qquad (5)$$

where

$$f(x) = \begin{cases} 0.0 & \text{if } |x| \leq F_{0\,th}, \\ \text{const.} & \text{otherwise.} \end{cases} \quad (6)$$

Difference cost reflects the F_0 difference between adjacent target segments and the selected phoneme units. It is defined in the concat subcost as

$$C_j^c(u_{i-1}, u_i) = |u'_{f_0\,i} - u_{f_0\,i}| \quad (7)$$

where

$$u'_{f_0\,i} = u_{f_0\,i-1} + t_{f_0\,i} - t_{f_0\,i-1}. \quad (8)$$

$t_{f_0\,i-1}$ and $t_{f_0\,i}$ denote target F_0 of the $(i-1)$–th and i–th segments, respectively. $u_{f_0\,i-1}$ and $u_{f_0\,i}$ denote the F_0 values of the $(i-1)$–th and i–th candidate phonemes, respectively. This cost is useful to keep a relative value of target F_0.

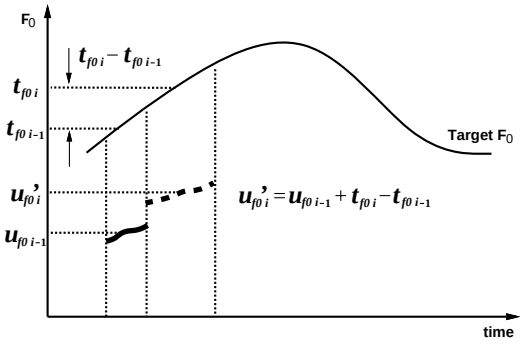

Figure 2: Difference cost.

4. EVALUATION

4.1. Experiments

A subjective hearing test was carried out to evaluate the effectiveness of the proposed subcosts for unit selection. Fifty Japanese newspaper sentences were used for producing synthetic speech through CHATR. Accents of the sentences were predicted automatically and corrected where necessary by hand. A Japanese female speech database was used to synthesize the speech. There were six subjects. Since the additional subcosts are all F_0 specific and may have side effects on other acoustic features, they classified the evaluated naturalness of (a) intonation and (b) continuity and clarity into five grades (from 'excellent' to 'bad'). The utterances were synthesized in five ways:

(1) using the original CHATR algorithm,

(2) using slope cost,

(3) using threshold cost (the threshold was set to 20 Hz from a preliminary examination),

(4) using difference cost, and

(5) using all three cost functions mentioned above simultaneously (all features).

The target F_0 was predicted based on J_ToBI [5][6],which is the Japanese version of ToBI system.

4.2. Results

Figure 3 illustrates the experimental results of the intonation evaluation. It shows the evaluation of 'bad' or 'poor' decreased about 20% and 'good' or 'excellent' increased about 10% by adopting each subcost. Furthermore, when all of these subcosts were adopted, the evaluation of 'bad' or 'poor' was reduced to about one-half of the original. This shows that these subcosts are effective in improving the intonation naturalness of synthetic utterances.

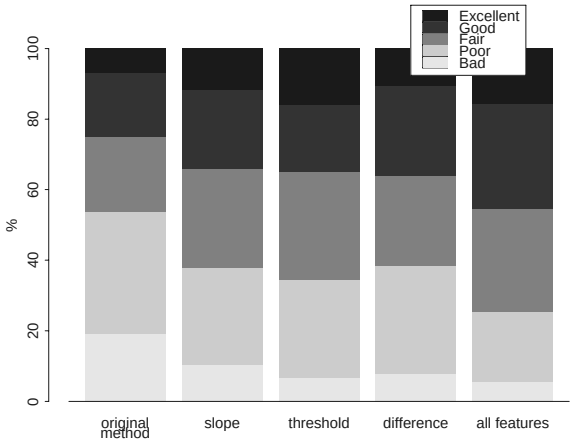

Figure 3: Evaluation of intonation.

Figure 4 illustrates the evaluation of continuity and clarity. It shows that all subcosts except slope cost degraded the continuity and clarity of synthetic speech compared with the original method. This is because the weight of other subcosts such as cepstrum and duration became relatively smaller by adding these subcosts, and the discontinuity between phonemes became larger. Slope cost, however, did not affect the continuity and clarity evaluation.

CHATR selects adjacent phonemes in the speech database when continuity between segments is dominant. This leads synthetic speech to be continuous, so the average number of adjacent

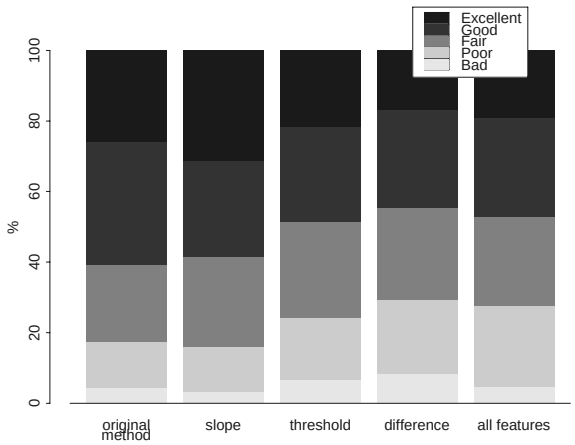

Figure 4: Evaluation of continuity and clarity.

phonemes in the speech database (average segment size) can be considered as a measure of continuity. Table 1 shows the average number of phonemes in a segment which are continuous in a speech database file. The average segment size in the utterances synthesized by the original algorithm and slope cost is larger than others, especially when compared with all features. This is consistent with the results of the subjective hearing test for continuity and clarity.

Table 1: Average speech segment size.

original method	slope	threshold	difference	all
1.7	1.7	1.6	1.5	1.3

Figure 5 illustrates the intonation and continuity–clarity evaluation results that were evaluated 'good' or 'excellent'. Applying all subcosts is best for improving naturalness of intonation, but it degrades the naturalness in terms of continuity and clarity. Since continuity and clarity distortion leads to a degradation in the naturalness of synthetic speech, we conclude that only slope cost should be applied for improving intonation naturalness at this stage.

5. CONCLUSIONS

This paper tested methods for improving the intonation naturalness of synthesized speech produced by the concatenative synthesis system CHATR. The intonation naturalness of synthesized speech was improved about 10% without distortion of continuity or clarity by considering the F_0 slope in unit selection. Setting a F_0 threshold for candidate phonemes

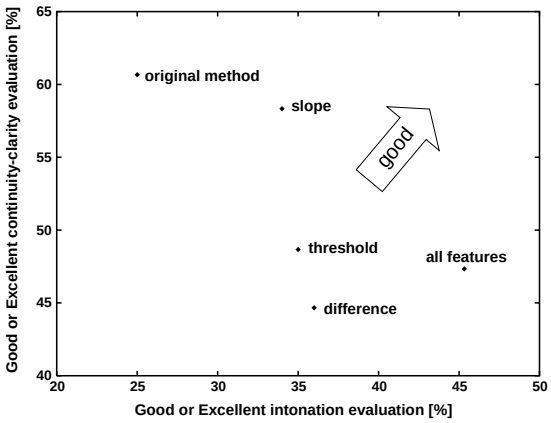

Figure 5: Intonation v.s. Continuity–Clarity evaluation classified as 'good' or 'excellent'.

and reflecting the F_0 difference between adjacent target segments and the candidate phonemes were both effective ways for improving intonation naturalness, but they also caused distortion of continuity and clarity in the synthesized speech. In order to improve intonation naturalness without causing distortion, future work should consider partial application of the above measures.

Acknowledgments

We would like to thank Dr. Nick Campbell for giving us the original idea for this study. We are also grateful to all of the members of Department 2 of ATR–ITL for their useful advice and cooperation in the evaluation tests.

REFERENCES

[1] N. Campbell. CHATR: A high-definition speech resequencing system. In *Proc. 3rd ASA/ASJ Joint Meeting*, pp. 1223–1228, Dec 1996.

[2] A. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proc. Eurospeech95*, pp. 581–584, Apr. 1995.

[3] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Speech Communication vol.9, 5/6*, pp. 453–467, 1990.

[4] D. Hirst and R. Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix, 15*, pp. 71–85, 1980.

[5] N. Campbell and J. Venditti. J-ToBI: an intonation labelling system for Japanese. In *Proc. Fall Meeting, Acoust. Soc. Jpn.*, pp. 317–318, Sept. 1995.

[6] J. Pierrehumbert and M. Beckman. *Japanese tone structure*. MIT Press, Massachusetts, 1988.