PROSODY–BASED UNIT–SELECTION FOR JAPANESE SPEECH SYNTHESIS

Ken Fujisawa and Nick Campbell

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

ABSTRACT

A corpus-based concatenative speech synthesis system using no signal processing can produce intelligible synthetic speech maintaining original voice characteristics. In such a concatenative system, it is very important to select appropriate waveform segments that are naturally close to the target prosody. But with limited size database it can sometimes be difficult to realize natural prosody.

This paper describes an approach to unit (waveform segment) selection for improving the intonation. We analyzed the pitch patterns of 503 sentences of read speech spoken by a Japanese female and obtained the F_0 range of natural prosody. Then we applied this restriction to the unit selection of the concatenative speech synthesizer. Through subjective experiments, we confirmed that this measure significantly improved the intonational naturalness of synthetic speech.

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 it preserves the voice characteristics of the original speaker. Prosodic features such as fundamental frequency(F_0) pattern and duration are used for unit selection by comparing them with the target features predicted for an input utterance. However, if suitable units are not found, the intonation of the synthetic speech can sound unnatural. This may be due to two reasons: one is the inappropriate prediction of the target F_0 pattern, and the other is inadequate unit selection. In a previous study [3], we proposed F_0 slope and other features to improve the prosody of CHATR. They improved the intonation of synthetic speech, but they tend to degrade the continuity naturalness. Therefore, we first analyzed the pitch patterns of 503 samples of read speech spoken by a Japanese female, and then used ToBI labels [4] to make a pitch pattern model for each accentual phrase in terms of (1) location of accentual phrase, (2) mora length and (3) accent position.

Next, the above pitch pattern model was introduced for the unit selection of CHATR. Phoneme-size candidate segments were selected for each accentual phrase by referring to the model pitch pattern, and then were pruned to minimize a concatenation cost.

2. CHATR

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 30 prosodic and phonetic factors, including duration, power, and F_0 at the center of each phoneme.

The concat cost $C^c(u_{i-1}, u_i)$ represents the distance between a selected unit and the adjacent 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

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

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

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

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

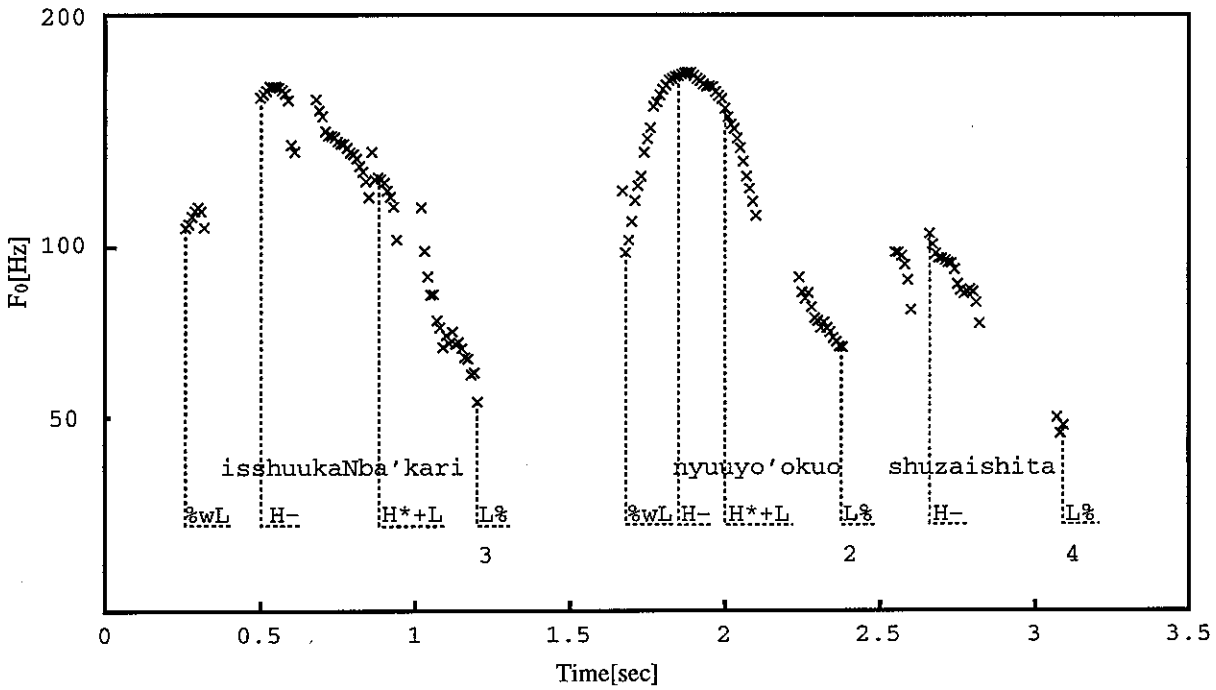

Figure 1: ToBI labeling example

Utterance: "isshuukaNbakari nyuuyookuo shuzaishita" (I collected news materials in New York for a week.) The '×' marks show F_0, which is derived from spoken speech. Word, Tone and BI Tiers are shown respectively below the F_0 pattern.

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.

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

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. JAPANESE TOBI LABELING

The ToBI(Tones and Break Indices) labeling system is used to represent the prosody for Japanese [5]. It consists of four tiers:

Word Tier romanised transcription of the words in the utterance.

Tone Tier pitch events in the F_0 contour.

BI (Break Index) Tier measure of the degree of association between two consecutive units.

Miscellaneous Tier other phenomena present in the speech signal that cannot be properly described, e.g., laughing, lip noise, etc.

Here, we describe the Tone and BI Tier, which are related to prosody.

Tone Tier

H*+L pitch accent marking.

H- phrasal tone marking the high F_0 of unaccented phrases.

L% final low boundary tone characterizing the accentual phrase in Japanese.

%L initial low boundary tone marked at the beginning of post-pausal phrase.

H% final high boundary tone marking.

BI Tier

0 junctures in fast speech phenomenon

1 word boundary

2 accentual phrase boundary

3 intonational phrasal boundary

4 end of an utterance

We use only BI2 – BI4 labels for the BI Tier in the current ToBI labeling. Figure 1 presents an example of ToBI labeling.

4. PROPOSED METHOD FOR UNIT SELECTION

Currently, CHATR predicts a target F_0 from accented romaji divided by break indices using a linear–regression model trained on ToBI labels. We propose a new method which combines prediction and selection in order to overcome some of the deficiencies of the present method. By modeling the F_0 characteristics of the accentual–phrase as a whole, we effect improvements in the synthesised intonation.

4.1. Speech database statistics

We analyzed the pitch patterns of 503 samples of read speech spoken by a Japanese female (FTK). All accentual phrases were categorized by (1) left and right BI of accentual phrase, (2) mora length, and (3) accent type with ToBI labels. The accentual phrase boundaries were given from the 'L%' label, and the accent types were given from the 'H*+L' label. Figure 2 illustrates the distribution of mora length in an accentual phrase from 503 utterances in the FTK speech database.

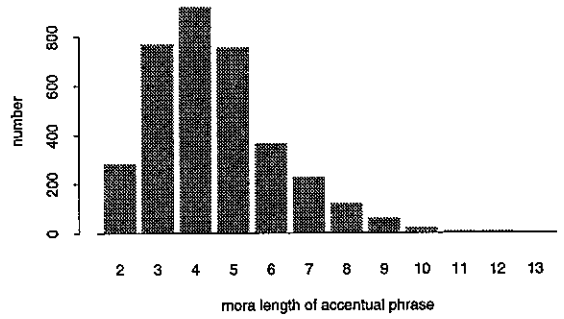

Figure 2: Mora length distribution in accentual phrase of 'FTK' speech database

For each category of accentual phrase, the following is calculated:

- average F_0 of each mora $(m_{l,t,p,n,i})$,
- average F_0 differences of adjacent morae $(d_{l,t,p,n,i})$, and
- standard deviation of F_0 differences of adjacent morae $(s_{l,t,p,n,i})$

here l is mora length of accentual phrase, t is accent type, p and n are left and right BI of related accentual phrase, and i represents the i–th mora in the accentual phrase. The accentual phrases that contain unvoiced vowels or double consonants were not used to create this F_0 range database.

4.2. Replacing target F_0

In order to improve the intonation prediction component, we implemented a whole-phrase model of F_0 control, in which the F_0 of each mora (or vowel) is assigned to $m_{l,t,p,n,i}$ if the proper accentual phrase category is found in the F_0 range database.

4.3. F_0 restriction in unit selection

We add the following subcost function to the concat cost to restrict the selected unit's F_0.

$$C_j^c(u_{i-1}, u_i) = \begin{cases} 0.0 & \text{if } u_i - u_{i-1} \leq d_{l,t,p,n,i} \pm \alpha s_{l,t,p,n,i} \\ \text{const.} & \text{otherwise.} \end{cases} \quad (5)$$

Here, u_{i-1} and u_i denote $(i-1)$–th and i–th unit's F_0, respectively. α is a variable that makes it possible to change the acceptable F_0 range. Const. is large enough to exclude the unit from synthesis. This subcost is considered only for vowels and 'N' (the vocalic nasal). Figure 3 illustrates this F_0 restriction cost. In the actual calculation, we used z–score of F_0 instead of the raw F_0 to eliminate dependency on the speaker's pitch range.

$$x_i^z = \frac{x_i - \overline{x}}{\sigma_x} \quad (6)$$

Here, σ_x denotes the standard deviation of speaker FTK's F_0, and $\overline{x}$ denotes the average F_0 of FTK. We can apply this subcost not only to FTK but also to other databases by using the z–score transfer. Incidentally, the F_0 range database is not sufficient for arbitrary input since it has too few data for accentual phrases which are more than 10 mora length. We confirmed from a preliminary subjective experiment that most of the intonational problems of synthetic speech can be categorized as either incorrect accent type or plural F0 falls within an accentual phrase. It is known that F_0 never rises twice (in other words, it never falls more than once) within an accentual phrase in Tokyo Japanese. If a corresponding accentual phrase pattern was not found in the F_0 range database, we applied the following default F_0 restriction:

- up to second mora in an accentual phrase, F_0 should rise unless the first or second mora is not accented.
- accented mora should have falling F_0.
- all mora after the accented one have gently falling F_0.

5. EVALUATION

5.1. Experiments

A subjective hearing test was carried out to evaluate the effectiveness of the proposed subcost for unit

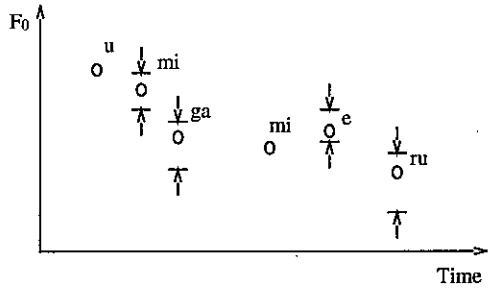

Figure 3: F_0 restriction at unit selection stage

Synthesis of an utterance: "u'miga 3 mie'ru" (The sea can be seen.) " ' " means accent, "3" means break index.

selection. Fifty Japanese sentences were generated by CHATR. The input for CHATR was accented romaji strings divided by break indices. Accents and break indices of the sentences were predicted automatically and corrected where necessary by hand. Japanese female (FTK) and male (MHN) speech databases were used to synthesize the speech. There were twelve subjects. The subjects listened to randomized pairs of speech samples, both synthesized by conventional CHATR and by the proposed method; the subjects selected the one that seemed to have the more natural intonation. Subjects were able to listen to the speech as many times as they liked.

5.2. Results

The results were evaluated by sentences, and if more than eight subjects evaluate one speech sample as prefered, it was regarded as better speech than the other. If only five to seven subjects evaluate one sample as better, it was regarded as even. Figure 4 illustrates the intonation evaluation results. Here, FTK and MHN denote the evaluation of unaltered CHATR and FTKnew and MHNnew denote the evaluation of CHATR using the proposed cost function at the unit selection stage. The results confirm that sentences generated using the new cost function can improve the intonational naturalness of synthetic speech. The improvement for speaker MHN is larger than that for FTK, even though the F_0 range database is derived from the FTK speech database. This may be because speaker MHN had a lower evaluation of intonation in the preliminary experiment [6].

Table 1: Evaluation result

Speaker	Better	Worse	Same
FTKnew	64 %	8 %	28 %
MHNnew	70 %	12 %	18 %

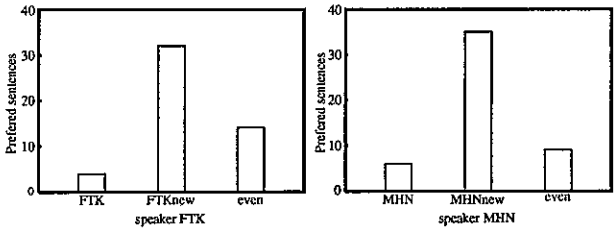

Figure 4: Intonation evaluation

6. CONCLUSIONS

This paper tested methods for improving the intonation naturalness of synthesized speech produced by the concatenative speech synthesis system CHATR. A Japanese female speech database was used to obtain the pitch ranges, which represented acceptable natural intonation. They were categorized by mora length, left and right break indices, and accent type of the accentual phrase. We applied this F_0 range knowledge to the unit selection of CHATR. Subjective hearing experiments confirmed that pitch range restriction at the unit selection stage can improve the intonational naturalness of CHATR. We also confirmed that the pitch range restriction obtained from a female speech database can be applied to the unit selection of a male database by using z–scored F_0 values.

Acknowledgments

We are grateful to all of the members of Department 2 of ATR–ITL for their useful advice and cooperation in the evaluations.

References

[1] N. Campbell. CHATR: A High-Definition Speech Re-Sequencing 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] K. Fujisawa, T. Hirai, and N. Higuchi. Use of pitch pattern improvement in the CHATR speech speech synthesis system. In *Proc. Eurospeech97*, pp. 2671–2674, Sep. 1997.

[4] N. Campbell. Autolabelling Japanese ToBI. In *Proc. of ICSLP*, pp. 2399–2402, Oct 1996.

[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] N. Campbell, Y. Itoh, W. Ding, and N. Higuchi. Factors affecting perceived quality and intelligibility in the CHATR concatenative speech synthesiser. In *Proc. Eurospeech97*, pp. 2635–2638, Sep. 1997.