Target Cost of F_0 Based on Polynomial Regression in Concatenative Speech Synthesis

Kei Fujii[*], **Hideki Kashioka**[*†] and **Nick Campbell**[*‡⋆]

[*] Nara Institute of Science and Technology
[†] ATR Spoken Language Translation Research Laboratories
[‡] ATR Human Information Science Laboratories
[⋆] JST/CREST Expressive Speech Processing
kei-fu@is.aist-nara.ac.jp, {hideki.kashioka, nick}@atr.co.jp

ABSTRACT

This paper proposes a target cost function for F_0 based on polynomial regression for use in concatenative speech synthesis. Polynomial regression is used to express the time series of F_0 continuously, and remove effects of microprosody. We conducted a perceptual experiment and confirmed that the proposed function provides a higher correlation with perceptual scores than does the conventionally used cost function.

1 INTRODUCTION

In concatenative speech synthesis, units are selected from a speech corpus based on minimisation of costs (by a weighted summation of target costs and concatenation costs) to obtain natural sounding synthetic speech [1]. Each unit is defined by a vector of features that include segmental characteristics, spectral features and prosodic features (duration, power and F_0), which are used for calculation of the costs.

In Japanese, accurate realisation of a pitch contour is required for the realization of lexical accents and the marking of interrogative form. In this paper, we focus upon improvements of the F_0 specification and calculation of the target cost for F_0 as a means to improve concatenative synthesis of Japanese speech. We believe, though, that the methods will be directly applicable for synthesis of other languages.

In conventional cost calculation, the F_0 of a unit is expressed as a sequence of N values representing the time series of F_0 of the unit. This method typically uses only a small number of values to represent the entire F_0 contour of the unit. For example, a value of $N = 1$ would typically represent the mean F_0 of the unit, $N = 2$ might be the start and end F_0 points of the unit. A value of $N = 3$ is common for a syllable-sized segment of speech.

Fujisawa et al. [4] reported that the naturalness of intonation in synthetic speech can be improved by making use of the F_0 slope derived from a linear regression as well as the average F_0 per unit for unit selection. We propose an improvement to their algorithm, whereby the slope contour of F_0 is obtained by use of a differentiating equation (3), to take account of non-linearities in the F_0 contour, and propose an improved target F_0 slope cost, derived as follows

$$C_t(Tgt_i, U_i) = [\int_s^e \{F_0'(Tgt_i, t) - F_0'(U_i, t)\}^2 dt]^d. \quad (1)$$

We propose the use of an expression which better represents the shape of the F_0 contour, as a step towards the realization of more high-fidelity synthetic speech. We therefore employ a polynomial regression in order to express the time series of the F_0 contour continuously, and calculate the cost as a function of the distance between the regression lines. The effects of microprosody are eliminated or reduced by using the polynomial regression [2], as described in section 3.

The proposed target cost function for F_0 is defined as the definite integral of the square of the difference between the polynomials of the candidate unit F_0 and the target F_0. This method of cost calculation increases the of F_0 in the unit-selection procedure without increasing the number of sub-costs in the overall unit selection process.

2 CONVENTIONAL F_0 TARGET COST

The target cost, which is calculated acording to the distance between a candidate unit and a target unit is defined as the weighted sum of subcosts [1]. The F_0 target cost that we focus on in this paper is one of the sub costs, and is defined as

$$C_t(Tgt_i, U_i) = \frac{1}{N}\sum_{j=1}^{N} |F_0(Tgt_i, j) - F_0(U_i, j)|, \quad (2)$$

where Tgt_i and U_i represent target unit and candidate unit respectively. N is the number of values representing the time series of the F_0 in each unit. $F_0(Tgt_i, j)$

ISBN 1-876346-48-5

indicates the jth F_0 value in Tgt_i. Similary, jth F_0 value in U_i is given by $F_0(U_i, j)$.

The method of selecting $F_0(U_i, *)$ from a time series of F_0 values in a unit U_i is represented schematically in figure 1. As figure 1 shows, $F_0(U_i, *)$ is chosen to divide U_i equally.

As shown by equation (2), the conventional target F_0 cost is defined as an average of the absolute values of the difference between a sequence of target F_0 representations and a sequence of candidate F_0 representations. i.e., the conventional method treats an F_0 contour as a discrete quantity, and when N is small, there is a possibility that the cost is based on a value by which only a part of the F_0 contour of a unit is considered.

For the present experiment, and in accordance with common practice, we measured F_0 on syllable-sized units (typically CV in Japanese), so that in many cases both consonant and vowel F_0 values are included in $F_0(U_i, *)$. Thus the conventional cost can be affected by microprosody occurring mainly from the influence of the consonants abutting each vowel. This can cause loss of accuracy in the cost function if it fails to represent the underlying (macroprosodic) pitch of the syllable, which we believe is closer to that of human perception.

To overcome this problem, we employed a microprosody removal process, as shown in section 3. The cost calculation method using polynomial regression is described in section 4.

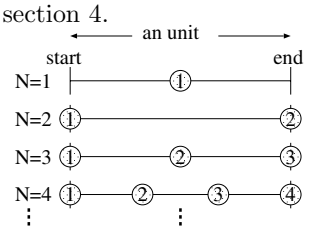

Figure 1: Selection of $F_0(U_i, *)$ corresponding to N in equation (2)

3 MICROPROSODY REMOVAL

Several techniques have been proposed in the literature for the removal of microprosody. Hirst et al. propose a method for modelling macroprosody automatically by fitting the contour with a quadratic spline function [3]. This method has been tested for several languages.

Fujisaki's model represents an F_0 contour by both a phrase component and an accent component. The parameters used in this model are obtained by applying an Analysis-by-Synthesis technique. F_0 extraction error and microprosody can cause degradation of the estimation accuracy, but Fujisaki and Narusawa have recently proposed a pre-processing stage to resolve this problem [2].

The pre-processing uses a third-order polynomial equation for smoothing and interpolation over intervals of voiceless consonants. It thus produces a representation of the F_0 contour which is continuous and differentiable. We employed this method for pre-processing our speech data before fitting a polynomial for the cost calculation as described in the next section.

4 TARGET F_0 COST BASED ON POLYNOMIAL REGRESSION

In this section, the proposed cost function is described. First, we describe feature extraction for the cost calculation, then the cost function itself.

4.1 EXPRESSION OF F_0 CONTOUR FOR PROPOSED COST FUNCTION

The F_0 contour of each unit is represented in the form of a polynomial. First, pre-processing [2] is applied to the estimated F_0 contour to remove and interpolate through microprosodic effects, then a fit m-order polynomial regression model to the resulting F_0 contour of each unit. For convenience, the time length of each unit is normalized to 1.

The F_0 contour of a unit U_i is expressed by

$$F_0(U_i, t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_m t^m, \quad (0 \le t \le 1), \tag{3}$$

where $\{a_0, \cdots, a_m\}$ are m-order polynomial regression coefficients of F_0 included in U_i. When m is set to 0, $F_0(U_i, t)$ has the same value as the mean of F_0. The case of $m = 1$ is identical to linear regression. In this paper, we investigated from $m = 0$ to 3.

Target F_0 is also expressed by the above equation. $\{b_0, \cdots, b_m\}$ denotes the polynomial regression coefficients of target F_0.

4.2 PROPOSED COST FUNCTION

The cost between the F_0 of a target unit(Tgt_i) and the F_0 of a candidate unit(U_i) is calculated by the following equation:

$$C_t(Tgt_i, U_i) = [\int_s^e \{F_0(Tgt_i, t) - F_0(U_i, t)\}^2 dt]^d \tag{4}$$

$$= \{\int_s^e (c_0 + c_1 t + c_2 t^2 + \cdots + c_m t^m)^2 dt\}^d, \tag{5}$$

where

$$c_n = a_n - b_n \quad (n = 0, \cdots, m), \tag{6}$$

d ... *adjustment constant*,

s, e ... *integral range* $(0 \le s < e \le 1)$.

As equation (4) shows, the proposed cost function is defined as a definite integral of the square of the difference between target F_0 and candidate unit F_0.

Although it may be undesirable from the point-of-view of perception to square the difference, we employ this operation to simplify the calculation. The constant d is adopted for adjust for any deterioration.

ISBN 1-876346-48-5

For the present experiment, the constants used in equation (4) were set to $d = 0.3$, $s = 0$ and $e = 1$ respectively.

5 EVALUATION EXPERIMENT

We performed a perceptual experiment to evaluate the proposed method described in the previous section, A cost function that has a higher correlation with the results of the perceptual evaluation is more desirable. Accordingly, we collected perceptual evaluation scores for a series of synthesised utterances. In following subsection, we describe the stimuli preparation, and the collection of perceptual evaluation scores.

5.1 STIMULI

Figure 2 depicts the design of the experiment. Subjects were presented with two sound samples, i.e. the natural speech sound of one short phrase, and the synthetic speech in which one syllable in the phrase was replaced by another equivalent syllable from a different phrase for comparison and evaluation. For example, in figure 2, the syllable /ko/ in the phrase /sonokotode/ is replaced to another syllable /ko/ from a different but phonetically similar context. Subjects listened to these two sound samples and evaluated the degree of distortion (if any) in the synthesised utterance.

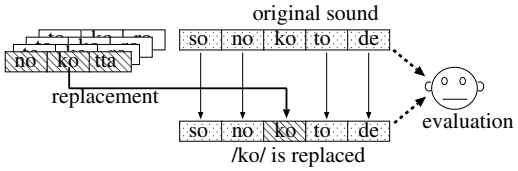

Figure 2: Graphical depiction of the experiment

There were two conditions on stimulus selection:

- All differences between the syllables, except in F_0, must be reduced to a minimum.
- The stimuli should have varing degrees of difference in F_0.

To satisfy the former condition, we extracted candidate syllables from the speech database by standard unit-selection criteria but with the weight of the target F_0 cost set to zero. We then selected a subset of these to use in the experiment by maximising the F_0 variation.

The procedure is shown below:

1. Set the target F_0 cost to zero.
2. For each unit U_i in the corpus, assume U_i to be the target, and perform unit-selection to obtain U_j.
3. Delete the overlapping combination.
4. For each combination (U_i, U_j), set N to 20 and compute the conventional F_0 cost as explained in section 2.

Because our speech data is Japanese, it has been labelled at the level of the mora. We extracted 11588 *morae* from a read-speech corpus of about 45 minutes duration (male speaker, 16 bit, 20 kHz sampling, 15518 *morae*) by the above procedure.

In order to select samples for use in the experiment, we classified these morae into three types; (a) CV where the C is voiced, (b) CV where the C is unvoiced, and (c) V (or the nasal mora) without a preceeding C.

Samples were then selected from each of the three categories by the following procedure:

1. Find the unit U_k which is the furthest from the average of the F_0 cost values.
2. Add U_k to the list of experiment samples.
3. Find the unit U_l which is furthest from all units in the list.
4. Add U_l to the list of experiment samples.
5. Repeat steps 3 and 4 until sufficient samples are found.

We collected 30 samples within each category for the experiment.

5.2 PERCEPTUAL EXPERIMENT

90 pairs of sounds were chosen according to the procedure in the previous section. This subsection explains the perceptual experiment which uses these samples.

Three subjects listened to these 90 pairs using headphones in a sound-damped room. The subjects were permitted to listen to the stimuli repeatedly if necessary until they were satisfied with their evaluation. The scores ranged from one (no difference can be perceived) to five (there is very large difference). Subjects were given practice with 30 training samples before the experiment proper began.

The average of the z-scores (normalisation to zero mean, unit variance) of the collected evaluation scores was treated as the ideal target F_0 cost.

6 RESULT

The correlations between the perceptual evaluation data collected as described in the previous section and each method of unit selection are displayed in table 1. The values in the column "num. of params." corresponds to the setting of N (which is $m+1$). Results for each class of mora are presented separately in tables 2, 3 and 4.

As we can see from the tables, the proposed cost function obtained a higher value of correlation with the perceptual scores than that of the conventional method in almost all cases.

ISBN 1-876346-48-5

num. of params.	conventional	proposed
1	0.676	0.791
2	0.531	0.819
3	0.633	0.823
4	0.685	0.819
20	0.728	

Table 1: Correlation of each method and perceptual evaluation values (all data)

num. of params.	conventional	proposed
1	0.812	0.819
2	0.436	0.834
3	0.636	0.834
4	0.705	0.834
20	0.752	

Table 2: Correlation of each method and perceptual evaluation values (solo vowel or nasal mora)

num. of params.	conventional	proposed
1	0.855	0.828
2	0.729	0.885
3	0.797	0.885
4	0.813	0.869
20	0.837	

Table 3: Correlation of each method and perceptual evaluation values (vowel with voiced consonant)

num. of params.	conventional	proposed
1	0.434	0.751
2	0.361	0.769
3	0.436	0.774
4	0.542	0.777
20	0.610	

Table 4: Correlation of each method and perceptual evaluation values (vowel with unvoiced consonant)

7 DISCUSSION

If we first examine results for the conventional method, we see that the conventional costs for $N = 2$ perform the worst in all cases. We assume that this degradation is caused as a result of taking F_0 measurements from both edges of the unit (see figure 1), where microprosody arising from the intervocalic consonants can be assumed to have most influence.

On the other hand, for voiced morae, we find better correlations for $N = 1$ in table 2 and 3, indicating that the influence of microprosody is smallest at the center of the mora. The correlation is not good for morae with unvoiced consonants (table 4), but when F_0 is estimated at the center of the vocalic portion, the correlation improves to 0.705. This indicates that the accuracy of the conventional cost function can vary significantly according to the position at which F_0 is calculated.

For the proposed cost function we see that the correlations with the perceptual scores vary considerably less than those of the conventional method when N or m is changed. In all cases, correlations for a higher-order N are greater than when N is set to 1, unlike the conventional method. Table 1 shows that N=3 gives the best performance, implying an improvement of the simpler linear model proposed by Fujisawa. Correlations for vowels following an unvoiced consonent (the most difficult case for F_0-based unit selection) also show increases with model order, further justifying the use of the proposed method for unit selection.

We conclude from the above that the suggested improvements are worthwhile, and we will now continue this work by testing our model on a large (100-hour) corpus of natural spontaneous speech, where, in contrast to the read-speech database, the F_0 variations are also used to signal paralinguistic as well as linguistic information to the listener.

8 CONCLUSION

In this paper, we proposed a target cost function for F_0 based on a polynomial regression for use in concatenative speech synthesis. It allows finer specification of the pitch contour by treating the time series of F_0 as a continuous quantity, and removing microprosody. We performed an experiment to investigate the correlation between the proposed method and a human perception-based evaluation score. The proposed function showed a higher correlation than the conventional cost function in the experiment.

ACKNOWLEDGMENTS

The authors would like to thank the Japan Science & Technology Corporation for support, and the students at the Applied Linguistics Laboratory of NAIST for their cooperation.

REFERENCES

[1] N. Campbell and A. W. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, Ed., pp. 279–292. Springer Verlag, 1995.

[2] H. Fujisaki and S. Narusawa, "Automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. Symposium on Prosody and Speech Processing*, pp. 133–138. 2002.

[3] D. J. Hirst, A. Di Cristo and R. Espesser, *Levels of representation and levels of analysis for the description of intonation systems*, (Ed.) Prosody: Theory and Experiment, Kluwer Academic Press (also available from http://citeseer.nj.nec.com/hirst00levels.html).

[4] K. Fujisawa, T. Hirai and N. Higuchi, "Use of pitch pattern improvement in speech resequencing system CHATR," *Proc. Autumn Meeting of The Acoustical Society of Japan*, pp. 219–220. 1997 (written in Japanese).