2007－ＳＬＰ－69（44）
2007／12／21

波形重畳型音声合成の韻律と音質のためのF0傾きモデルの利用

立花　隆輝[†]　　長野　　徹[†]　　西村　雅史[†]

† 日本アイ・ビー・エム 東京基礎研究所　〒242–8502 神奈川県大和市下鶴間 1623–14
E-mail: †{ryuki,tohru3,nisimura}@jp.ibm.com

あらまし　波形重畳型や波形接続型の音声合成では、音声素片の接続部分で基本周波数の一貫性が失われることがある。日本語は高低アクセント言語であるため、これは訛りやアクセント型の誤りとして認識される問題となる。特にデータベースサイズが限られているときや、音声コーパス設計時に考慮していた想定ドメイン以外の文章でそのような問題が顕著になりやすい。本論文では、まずF0傾きモデルを用い音声素片のF0の一貫性を向上する。さらにF0傾きモデルも利用したF0修正量探索により、適切な素片が見つからない場合にも一貫性の回復を行う。これにより入力文章の想定ドメイン内外に関わらず、音声素片の音質を活かしつつ正しい高低アクセントで合成できるようになる。

キーワード　波形重畳型音声合成、韻律、基本周波数、アクセント

F0 Gradient Model for Acoustic Quality and F0 Consistency of Concatenative TTS

Ryuki TACHIBANA[†], Tohru NAGANO[†], and Masafumi NISHIMURA[†]

† Tokyo Research Lab., IBM Japan　Simotsuruma 1623–14, Yamato-shi, Kanagawa-ken, 242–8502 Japan
E-mail: †{ryuki,tohru3,nisimura}@jp.ibm.com

Abstract　A problem with concatenative text-to-speech synthesis is that it sometimes fails to preserve the appropriate consistency in the F0 contours at the concatenation points of the speech segments. Since Japanese is a pitch accent language, listeners perceive inconsistency in F0 contours as strange accents or wrong accent nuclei. Such problems occur more frequently when the database size is limited or when synthesizing voices for texts in new application domains. In this paper, we propose an F0 gradient model and F0 adjustment to select consistent speech segments and to restore the consistency by adjusting the F0 values only where necessary. This makes it possible to generate synthetic voices with correct pitch accents while taking advantages of the acoustic quality of the recorded speech segments even in new application domains.

Key words　Concatenative Text-to-Speech Synthesis, Prosody, Fundamental Frequency, Pitch Accent

1. Introduction

Concatenative text-to-speech (CTTS), also called unit selection, is one of the major approaches for TTS. When a system of this kind successfully finds a proper sequence of speech segments in its segment database (DB), it produces a synthetic voice with an acoustic quality that is almost as natural and lively as a human voice. One problem with the approach is that it requires a large DB. To tackle this problem, the number of candidate segments is often reduced by using criteria to preselect certain segments before the segment search. However, such preselection tends to cause inconsistency in the fundamental frequency (F0) contours of the selected speech segments. The concatenative approach selects speech segments whose prosodic parameter values are close to the target prosodic parameter values as estimated with prosody models. Since the selected segments are a compromise for multiple factors such as F0, duration, and energy, the concatenation of the selected segments sometimes lacks the necessary consistency in the resulting F0 contour. For example, for Japanese, a small F0 gap created by F0 errors may be perceived as an incorrect accent nucleus by listeners, or a necessary accent nucleus may be lost because of the F0 errors. Such problems occur more frequently when synthesizing voices for out-of-domain texts, because more frequent concatenation is required compared to in-domain texts. An in-domain text is a text in a domain for which a considerable number of recorded human voices are in the DB. An out-of-domain text is a text in other domains. A naïve method to deal with this problem is to force the synthetic voices to have the target F0 values by modifying the waveforms of the selected speech segments using pitch-synchronous overlap-

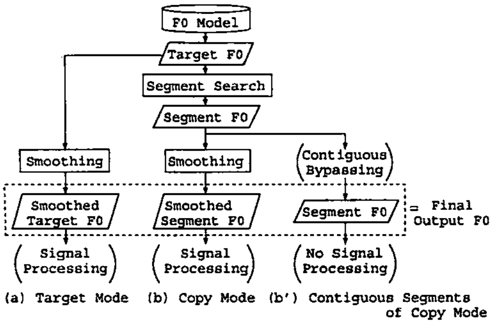


Figure 1 F0 handling of the conventional method.

and-add (PSOLA). However, this interferes with the merits of the approach, by reducing the liveliness and the acoustic quality of the synthetic voices.

How can we achieve a good balance between the acoustic quality and the prosodic consistency of synthetic voices with a concatenative method? In this paper, we propose a method to do this. The two main ideas of the proposed method are an F0 gradient model and F0 adjustment. The F0 gradient model is a stochastic model for modeling the consistency of the F0 contours using their gradients, and the model is used both in the segment search and the F0 adjustment. The F0 adjustment is an additional search process to search for a sequence of proper amounts of F0 modification that restore consistency in the final F0 contour. The effectiveness of the proposed method for the acoustic quality and Japanese pitch accents is shown by experimental results with subjective listening tests.

2. Conventional Concatenative Method

In this section, we describe a conventional method [1] from which we develop the proposed method. Due to limitations of space, we focus on how the method handles F0 and on the problems related to F0 treatments. Since other concatenative methods have some common features with this method, we consider the problems are also more or less similar for other concatenative methods.

2.1 F0 Handling

The minimum unit of concatenation is a sub-phoneme. The segment search searches for the sub-phoneme sequence with the minimum cost. The cost consists roughly of a target cost and a concatenation cost. The target cost is the weighted summation of an F0 cost, a duration cost, and an energy cost. These costs are penalties for the errors in the prosodic parameter values of the segments compared to the target prosodic parameter values. For example, the F0 cost for a speech segment is a penalty for the difference between the target F0 value and the segment F0 value (the F0 value the speech segment originally has). The concatenation cost is the weighted summation of a spectral continuity cost and an F0 transition cost. These costs are penalties for the gaps between each consecutive pair of the segments.

The target F0 values are predicted by using a decision tree. The set of F0-prediction features includes text-based information such as the pitch accent of the current syllable and the part-of-speech of the current word. A set of 14 features per syllable is extracted over a context window of five syllables consisting of the current syllable plus the two preceding and two following syllables. For each node of the tree, the average of the F0 distribution of the training data is used as the target F0 value for the node at run-time.

There are two basic options for handling the F0 values of the output speech. The first option (*Target*) is to use the smoothed target F0 values ignoring the segment F0 values (Fig. 1a). The other option (*Copy*) is to use the F0 values of the selected segments after removing the warbling by smoothing the F0 values (Fig. 1b). Though the speech signal is processed to match the resulting F0 contour in *either* case, *Copy* produces synthetic voices with a better acoustic quality because of the smaller amount of signal processing.

Smoothing sometimes damages the acoustic quality and the liveliness of the original voice. To preserve the original quality whenever possible, the method bypasses signal processing completely for sufficiently long sequences of segments which were contiguous in the original corpus. This mechanism (*Contiguous Bypassing*) is usually used in a combination with *Copy* (Fig. 1b').

The DB size can be reduced by eliminating the segments least likely to be used. This procedure (*Preselection*) is performed by collecting usage statistics of the segments [2]. The method synthesizes a large number of sentences, counts the number of times each speech segment is used, and then removes the segments with small counts. A DB reduction technique based on usage statistics has also been proposed by [3].

2.2 Problems

There are various causes of F0 contour inconsistency: (1) the F0 model may predict improper F0 values, (2) there may be no speech segments having F0 values close to the target F0 values, or (3) the varying F0 errors before and after concatenation points may produce unintended effects on the full F0 contour. The concatenation points have the highest probabilities of causing F0 contour inconsistency when *Copy* is used. If the DB is sufficiently large, it is likely that long contiguous speech segments having proper F0 values will be found, based on the target F0 values and the context information such as pitch accent values and phoneme identities. Therefore, a combination of *Copy* and *Contiguous Bypassing* seems to be the best choice for handling the F0 contours. However, when *Preselection* is used, the best choice depends on the size of the preselected DB, the domains of the application and the recorded voices, and listeners' preferences for synthetic voices. With aggressive *Preselection*, frequent concatenations are unavoidable. While the combination of *Copy* and *Contiguous Bypassing* may still produce good acoustic quality for in-domain texts, it may cause unstable F0 contours for out-of-domain texts. Use of *Target* would produce rather more stable F0 contours. However, the liveliness and the acoustic quality for the in-domain texts would be reduced.

3. Method

In this section, after the main ideas of the proposed method are briefly introduced, we describe the training procedure and the run-time processes.

3.1 Main Ideas

3.1.1 F0 Gradient Model

Because the human auditory system is more sensitive to frequency changes than to absolute frequencies, we use a GMM-based (Gaussian-Mixture-Model-based) F0 gradient

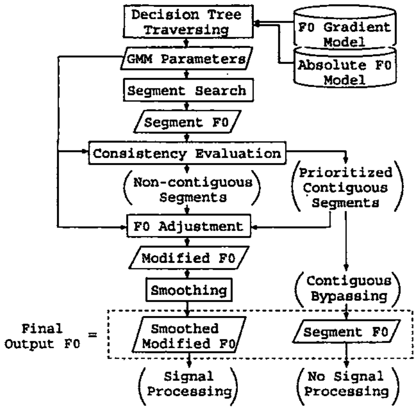


Figure 2 The process flow of the proposed method.

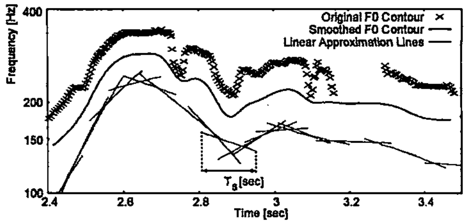


Figure 3 An example of a smoothed F0 contour and linear approximation lines.

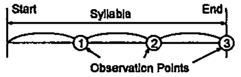


Figure 4 Three observation points for a syllable.

model to evaluate the consistency of the F0 contours. The model makes it possible to evaluate all of the segments in an interval for a fixed duration by linearly approximating the F0 values in the interval. Meanwhile, as long as consistency of F0 gradients is preserved, being close to pitch target values is unimportant. Therefore, we also use a GMM-based absolute F0 model, instead of target F0 values, to allow a wide range of absolute F0 values.

3.1.2 F0 Adjustment

The proposed method modifies the segment F0 values only where necessary, so that it is able to preserve high acoustic quality for in-domain texts and so that it is able to generate accurate F0 contours even for out-of-domain texts. To determine the proper F0 modifications, in addition to the existing segment search, we use a second search, called F0 adjustment, to search for the sequence of F0 modifications that minimizes the F0-related costs. In addition, the *Contiguous Bypassing* mechanism is also modified. Contiguous segments may not have the correct F0 contour in the context of the text to be synthesized. Hence, the method checks the F0 contours of the contiguous segments by using the F0 gradient model. Only the contiguous segments having appropriate consistencies (*Prioritized Contiguous Segment*) are given special priorities in the second search and are then used for *Contiguous Bypassing*.

3.2 Training Procedure

The F0 gradient model and the absolute F0 model are trained in the following procedure.

3.2.1 Smoothing

We smooth the F0 contours of the narrator's voices by convolving a Gaussian function to the original contours before collecting the training data. This is because overly detailed F0 changes are only noise for the models. Linear interpolation before smoothing can fill in missing F0 values for devocalized regions. An example of smoothing is shown in Fig. 3. The smoothed contour was shifted downward by 50 Hz for better visualization.

3.2.2 Observation

We check the absolute F0s and F0 gradients at three equally-spaced points for each syllable(注1) (as illustrated in Fig. 4). For an F0 gradient, we linearly approximate the logarithmic values of the smoothed F0s in the interval for a fixed duration (T_s seconds) ahead of each observation point. Examples of the linear approximation lines are shown in Fig. 3. These lines are shifted downward by 100 Hz for better visualization. We do not use the intercepts of these lines, but only the gradients of the lines are used for the model.

3.2.3 Decision Tree Training

The absolute F0 model and the F0 gradient model are trained in very similar manners. The input features are the 70 features of the conventional F0 model plus an additional feature indicating the observation point within the syllable. The values of 1, 2, and 3 correspond to the points that are 1/3, 2/3, and 3/3 from the beginning of syllable. Both of these models predict a scalar real value (an absolute F0 value or an F0 gradient). After the regression trees are trained, we build a GMM for each of the nodes of the trees, to model the distribution of the output features of that node.

3.3 Run-time Procedure

The proposed run-time procedure is different from the conventional method primarily in the following two ways: (1) it performs F0 adjustment after the speech segments are selected, and (2) both the segment search and the F0 adjustment use the new GMM-based likelihood costs for the F0-related costs. The processing steps related to F0 handling are as follows (and illustrated in Fig. 2).

3.3.1 Decision Tree Traversing

First, this step collects the input feature vectors for the decision trees based on the context information obtained by analyzing the input text. Then, by traversing the decision trees with the feature vectors, it obtains for later use a set of GMM parameters for each of the three observation points of each syllable.

3.3.2 Segment Search

The segment search searches for the best segment sequence with the minimal cost. The cost now consists of the three new F0-related costs as well as the other costs, such as the duration cost and the spectral continuity cost. Before calculating the F0-related costs for a segment, the closest F0 observation point within the syllable is determined for the segment. The three new costs are:

(1) The absolute F0 cost

We calculate the likelihood of the logarithmic F0 value, $\tilde{f}_{c,i}$,

(注1) : Actually mora in the case of Japanese.

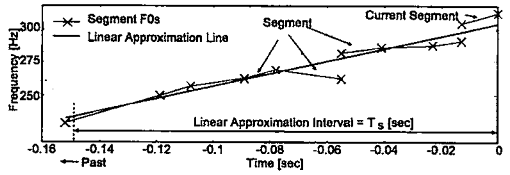


Figure 5 An example of segment F0s and an approximation line.

at the center of the current segment, where i has a value ranging from 1 to the number of segments, N_s, in the synthetic voice. The likelihood $Pr(\tilde{f}_{c,i})$ of the absolute F0 $\tilde{f}_{c,i}$ is then calculated by using the GMM parameters of the closest F0 observation point. The absolute F0 cost is finally calculated as $C_{a,i} = -w_a \log\left(Pr(\tilde{f}_{c,i})\right)$, where w_a is a given constant weight.

(2) The F0 gradient cost

To calculate this cost, we calculate the F0 gradient in the last T_s-second period of the current segment sequence. First, we take the logarithmic F0 values of the starting points and the ending points of the segments in the interval. By combining them with the durations of the segments, we can obtain the coordinates (x_j[sec], y_j[$log\ Hz$]) of these points. Then we find the line that best approximates the points (as shown in the example in Fig. 5). The gradient and the intercept of the line are g_i and s_i, respectively. The likelihood $Pr(g_i)$ for the pitch gradient g_i is calculated by using the GMM parameters for the gradient model. The F0 gradient cost is finally calculated as $C_{g,i} = -w_g \log\left(Pr(g_i)\right)$, where w_g is a given constant weight.

(3) The linear approximation error cost

We calculate this cost as

$$C_{f,i} = w_f \sqrt{\frac{1}{Nf}\sum_{j=1}^{Nf}\{y_j - (g_i x_j + s_i)\}^2}, \qquad (1)$$

where w_f is a given constant weight and Nf is the number of points to be approximated in the period. The first reason we use this cost is that the F0 gradient cost becomes invalid if the linear approximation error is too large. The second reason is that the change of the segment F0s in the short intervals such as T_s seconds should be smooth to allow for linear approximation.

3.3.3 Consistency Evaluation of Contiguous Segments

The *Contiguous Bypassing* mechanism of the existing method cancels the signal processing for long segment chunks when sufficiently long chunks are found in the DB. However, the chunks may not have the required consistency in the current context. If the chunks do not have sufficient consistency, we should also do signal processing for these chunks. If the chunks are adequately consistent, then we should leave the chunks unchanged and modify the neighboring segments to make best use of those chunks. For this analysis, we compare the F0 gradient costs of the contiguous segments with a given threshold. Only the contiguous segments whose costs are below the threshold, the *Prioritized Contiguous Segments*, are used in the following steps and to cancel the signal processing for these segments.

3.3.4 F0 Adjustment

This step searches for the best sequence of modifications for the segment F0 values. The i-th modification amount, m_i, is added both to the ending F0 of the $(i-1)$-th segment and the starting F0 of the i-th segment per the following equations (and as illustrated in Fig. 6):

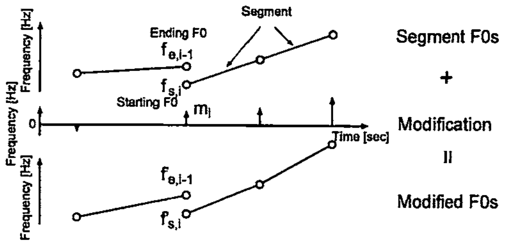


Figure 6 An F0 modification value is added to the ending F0 of the preceding segment and the starting F0 of the following segment.

$$f'_{e,i-1} = f_{e,i-1} + m_i \qquad (2)$$
$$f'_{s,i} = f_{s,i} + m_i, \qquad (3)$$

where i has a value from 1 to $N_s + 1$ and $f_{s,i}$ and $f_{e,i}$ are the F0 values at the starting point and the ending point of the i-th segment, respectively. $f'_{s,i}$ and $f'_{e,i}$ are the modified starting F0 and the modified ending F0, respectively. The value of m_i is chosen from a discrete candidate list (for example, $-100, -90, .., \pm 0, .., +90, +100[Hz]$). The best modification sequence $(m_1, .., m_{N_s+1})$ with the minimum cost is searched for by using a Viterbi algorithm. The cost of the modification sequence C_t evaluated at the i-th segment is the summation of C_a, C_g, C_f, and the modification cost C_m as $C_{t,i} = C_{a,i} + C_{g,i} + C_{f,i} + C_{m,i}$. We ignore the other costs such as the duration cost that are irrelevant to the F0 values. The modification cost is calculated by

$$C_{m,i} = \begin{cases} w_p m_i^2 & (\textit{Prioritized Segments}) \\ w_m m_i^2 & (\textit{otherwise}) \end{cases}, \qquad (4)$$

where w_m and w_p are given constant weights. When the i-th segment is a *Prioritized Contiguous Segment*, the special weight w_p with a huge value is used to block modification of the F0 values of the segments.

3.3.5 Final F0 Smoothing

This step determines the final F0 values of the segments by smoothing the modified F0 values. The *Prioritized Contiguous Segments* are treated as exception without smoothing by the *Contiguous Bypassing* mechanism.

4. Experiments and Results

We conducted subjective listening tests to investigate the effectiveness of the method, assessing both the acoustic quality and the pitch accent naturalness in a situation where the DB size is limited. We compared the following three methods: (*Copy*) for the conventional method with the *Copy* mode and the *Contiguous Bypassing* mechanism, (*Target*) for the conventional method with the *Target* mode, and (*Proposed*) for the proposed method. We used a speech corpus from a professional female narrator for building a TTS data set. Table 1 shows the statistics of the data set. The corpus contains readings of news, weather forecasts, traffic information, etc. We built an absolute F0 model with 164 nodes and an F0 gradient model with 267 nodes. Based on informal preliminary experiments, we decided to use 0.15 seconds as the

Table 1 The DB size before and after *Preselection.*

Data set	# of segments	Total duration
Before *Preselection*	1,185,329	10 hours
After *Preselection*	103,725	0.9 hours

Table 2 The results of the subjective listening tests.

	Pitch	Acoustic	RMSM
	Accent	Quality	[Hz]
Copy	3.39	3.33	12.6
Target	3.59	3.01	31.1
Proposed	4.08	3.20	18.0

duration of each linear approximation window (T_s). A GMM with 4 components was trained for each node. The conventional method used an F0 tree with 474 nodes to each of which an average F0 value was assigned.

To separately clarify the effects of the method on in-domain texts and out-of-domain texts, we assumed the preselected system was dedicated to traffic information and route navigation, which are included among the major domains in the original recorded DB. For this purpose, we performed preselection exclusively biased on the application domains. We collected usage counts for the segments by synthesizing the texts of the 8,851 sentences of the corpus. We multiplied the usage counts for 1,000 sentences in the application domains by 10 to preselect many segments in the domains, while we did not multiply the usage counts for the other 7,851 sentences with any weight. To make the experiments fair, the usage counts for both of the existing methods and the proposed method were collected and summed. Based on the summed usage counts, we built a preselected DB, which was used for both of the existing methods and the proposed method in the experiments. We noticed the segment search sometimes selected segments whose durations were seriously different from the duration targets. We used time-scale modification for these outliers in the run-time instead of manually removing such outliers from the DB.

4.1 Test Design

We generated synthetic voices for 100 in-domain sentences and 100 out-of-domain sentences by using each method. The in-domain sentences are from the application domains of traffic information or route navigation. The out-of-domain sentences are the first 100 sentences of the phonetically balanced ATR 503 sentences. These sentences were not included in the training data. We manually corrected the text processing output for the sentences to exclude the effects of the text processing accuracy.

The tests were 5-grade MOS (Mean Opinion Score) listening tests. A total of 15 subjects participated in the tests. The subjects evaluated each synthetic voice using two criteria, the naturalness of the pitch accents and the acoustic quality, by selecting answers from 5 (Very Natural or Very Good) to 1 (Very Unnatural or Very Bad). Each subject listened to 10 randomly chosen sets of synthetic voices. One set consists of synthetic voices generated with all three methods for the same sentence. The order of the methods in each set was random and we did not inform the subjects which system generated which voice. The first 5 sets of each subject were randomly chosen from the 100 out-of-domain sentences. The other 5 sets were randomly chosen from the 100 in-domain sentences. Since a sentence is too long to assess, each test sentence was divided into intonational phrases. We asked each subject to evaluate each of the phrases. Very short intonational phrases (< 1.5 seconds) were combined for the convenience of the subjects.

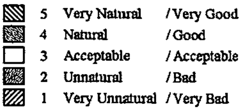
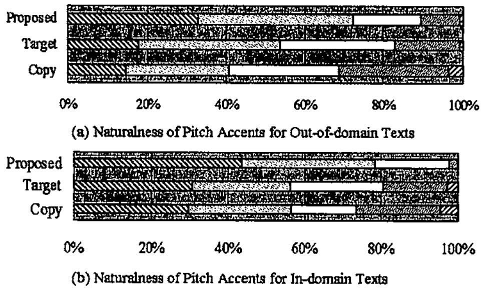


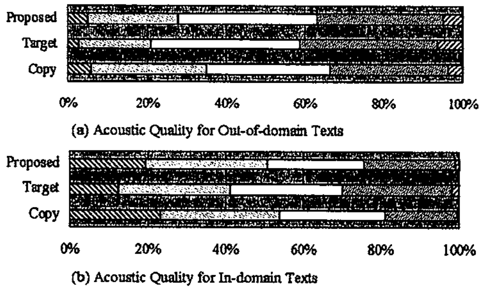
Figure 7 Graph legends for Fig. 8 and Fig. 9.



(a) Naturalness of Pitch Accents for Out-of-domain Texts

(b) Naturalness of Pitch Accents for In-domain Texts

Figure 8 Experimental results of accent naturalness.



(a) Acoustic Quality for Out-of-domain Texts

(b) Acoustic Quality for In-domain Texts

Figure 9 Experimental results of acoustic quality.

4.2 Results

Table 2, Fig. 8, and Fig. 9 show the experimental results. Fig. 8 shows the distribution of the assessments of the pitch accent naturalness for out-of-domain texts and in-domain texts. Fig. 9 shows the distribution of the assessments of the acoustic quality. Table 2 shows the average scores for the compared systems to give a brief overview of the results. In addition, we also show the values of *Root Mean Square Modification (RMSM)* in the table. An RMSM is calculated by

$$RMSM = \sqrt{\frac{1}{N_t}\sum^{N_t}\left(\frac{1}{N_s}\sum_{i=1}^{N_s} m_{f,i}^2\right)}, \qquad (5)$$

where N_t is the number of test sentences, N_s is the number of segments in the sentence, and $m_{f,i}$ is the difference between the original segment F0 of the i-th segment and the final F0 value. The RMSM can be considered as an indicator of the average modification amount. A large RMSM value indicates much signal processing and poor acoustic quality.

Fig. 8 shows that *Proposed* performed better than the other two methods both for the in-domain texts and the out-of-domain texts. *Copy* was the worst method in the pitch accent naturalness. The reason of the inadequate naturalness of *Target* was that predicting the absolute F0 values individually for each mora was not suitable for producing natural pitch accents of Japanese.

We can see in Fig. 9 that the acoustic quality of *Proposed* were better than *Target* though *Copy* was best in the acoustic quality both for the in-domain texts and the out-of-domain texts. We can see this rank order also in the RMSM values shown in Table 2.

You may wonder why the quality of *Target* was not far worse than that of *Copy* despite the different RMSM values. The acoustic quality scores of the three methods were generally poor. This was because the major source of the poor scores was discontinuity rather than signal processing, and because the small size of the preselected DB commonly used for the three methods caused the discontinuity.

5. Related Work

Some F0 modeling methods [4]~[6] in this field model distributions of delta F0 values as well as distributions of absolute F0 values. An advantage of our F0 gradient model over the delta F0 models is that the F0 gradient model takes all of the segments within the linear approximation interval into consideration, while the delta models compare the F0 values only with a few specific points in the past. In addition, if we try to use a delta F0 model for calculating the F0 likelihoods of sub-phonemes, the wide variety of the durations of sub-phonemes makes it difficult to choose which sub-phoneme from the past we should use for reference F0 values for the delta F0. In contrast, the F0 gradient model uses a fixed duration (T_s) regardless of the sub-phoneme durations.

For F0 treatments, leaving segment F0s unmodified is currently the most common treatment [7]. The main reason to avoid signal processing is to preserve the acoustic quality of the original voices. It was reported by [7] that signal processing is harmful to the acoustic quality rather than helpful to the prosody when there is more than 2 hours of speech data. However, there are pressures to reduce the size of the DB to less than 2 hours. In addition, [8] reported that intonation was still the largest problem for their system with a 2.5-GB uncompressed DB. For these reasons, we consider the naïve use of segment F0s to be sometimes problematic. In contrast, [9] modified the selected segments to the target F0 values. While that method has advantages in producing stable and smooth synthetic voices, it is difficult for the method to produce lively synthetic voices with the good acoustic quality found in unmodified segments.

To fix the pitch accents of a concatenative method, the methods [10], [11] use a support vector machine (SVM) to assess the correctness of the pitch accents. The SVM is trained with a set of training data labeled with "correct" or "wrong". After judging the correctness, [10] modifies the segments in the "wrong" phrases to have the target F0s, while [11] alters the "wrong" segment sequences to "correct" ones and concatenates the "correct" segments without modification. Though we have a similar objective, the largest difference is that our gradient model is not modeling the correctness of pitch accents, while their SVMs are trained to assess correctness. The main reason our model has the effect of producing correct pitch accents is because the narrator was originally producing correct pitch accents. However, since our models are not trained exclusively for modeling correct pitch accents, they should also be useful for reproduction of the narrator's speaking habits. In addition, our method searches for the sequence of proper amounts of F0 modifications instead of just using the target F0 values to minimize the modification amounts.

6. Conclusion

In this paper, we showed the effectiveness of the proposed method for the acoustic quality and the pitch accent naturalness by the experiments conducted separately for in-domain texts and out-of-domain texts.

Our future work includes additional subjective listening tests to clarify the contributions of each sub-mechanism of the proposed method. In addition, we also need to assess the effectiveness of the method for reproduction of the narrator's speaking habits.

References

[1] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1099–1108, 2006.

[2] W. Hamza and R. Donovan, "Data-Driven Segment Preselection in the IBM Trainable Speech Synthesis System," in *Proc. ICSLP*, September 2002, pp. 2609–2612.

[3] P. Rutten, M. Aylett, J. Fackrell, and P. Taylor, "A Statistically Motivated Database Pruning Technique for Unit Selection Synthesis," in *Proc. ICSLP*, September 2002, pp. 125–128.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," *IEICE Trans. Information and Systems (Japanese Edition)*, vol. J83-D-II, no. 11, pp. 2099–2107, November 2000 (in Japanese).

[5] X. Ma, W. Zhang, W. Zhu, Q. Shi, and L. Jin, "Probability Based Prosody Model For Unit Selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, pp. I–649–I–652.

[6] S. Shechtman, "Maximum-Likelihood Dynamic Intonation Model for Concatenative Text-to-Speech System," in *Proc. 6th ISCA Speech Synthesis Workshop*, August 2007, pp. 234–239.

[7] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A New TTS from ATR Based on Corpus-Based Technologies," in *Proc. 5th ISCA Speech Synthesis Workshop*, June 2004, pp. 179–184.

[8] K. Mano, H. Mizuno, H. Nakajima, H. Asano, M. Isogai, M. Hasebe, and A. Yoshida, "Development of a corpus-based concatenative text-to-speech synthesis system "Cralinet" for contact center services," in *Proc. Fall Meeting of Acoustic Society of Japan*, September 2004 (in Japanese), pp. 347–348.

[9] S. Buchholz, N. Braunschweiler, M. Morita, and G. Webster, "The Toshiba entry for the 2007 Blizzard Challenge," in *Proc. Blizzard Challenge 2007 Workshop*, August 2007.

[10] M. Hasebe and H. Mizuno, "Segment selection in corpus-based speech synthesis," in *Proc. Spring Meeting of Acoustic Society of Japan*, March 2004 (in Japanese), pp. 215–216.

[11] A. Yoshida, H. Mizuno, and K. Mano, "A unit reselection method based on accent evaluation for concatenative speech synthesis," in *IEICE Technical Reports*, August 2006, vol. SP2006-37, pp. 12–18.