論文 / 著書情報
Article / Book Information

Title	An Efficient Prosody Adaptation Method and Its Application to HMM-based Speech Synthesis
Author	Hosana Kamiyama, Takahiro Shinozaki, Koji Iwano, Sadaoki Furui
Journal/Book name	Proc. of the Second APSIPA Annual Summit and Conference, , , pp. 82-85
Issue date	2010, 12

An Efficient Prosody Adaptation Method and Its Application to HMM-based Speech Synthesis

Hosana Kamiyama *, Takahiro Shinozaki *, Koji Iwano† and Sadaoki Furui *
*Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: {khosana, shinot, furui} @furui.cs.titech.ac.jp Tel/Fax: +81-3-5734-3481
†Tokyo City University, 3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama, 224-8551 Japan
E-mail: iwano@tcu.ac.jp Tel/Fax: +81-45-910-2598/2599

Abstract— **This paper proposes an efficient prosody (F_0 contour and phoneme duration) adaptation method using a limited amount of utterances by a target speaker for speech synthesis. The proposed method uses dummy-variable regression models and only bias terms are adapted. Subjective evaluation results show that 5 sentences are enough for achieving natural and personalized prosody which is almost equivalent to that achieved using models trained by 470 sentences for each speaker. The proposed prosody adaptation method was combined with cepstral feature adaptation in the framework of HMM-based speech synthesis. Since the cepstral feature adaptation needs 20 sentences, 20 sentences are enough for adapting all the features to produce natural and personalized synthesis speech.**

I. INTRODUCTION

Recently, HMM-based speech synthesis has been actively studied. In various applications of speech synthesis, it is necessary to be able to produce synthesized speech with various individuality as well as naturalness. One of the advantages of the HMM-based speech synthesis in comparison with the waveform concatenation method is that HMMs can be easily adapted to a target speaker using a limited amount of utterances. Not only the cepstral features but also prosodic features, such as fundamental frequency (F_0) and phoneme duration, can be simultaneously modeled by HMMs and they can be simultaneously adapted using utterances by a target speaker [1].

In our framework, cepstral features are modeled by HMMs, but prosodic features are modeled separately by a dummy-variable regression model [2]. The system overview is shown in Fig. 1. An advantage of such separate modeling scheme is that it increases flexibility of model modification and adaptation. For example, a polyglot TTS (text-to-speech) system which produces multiple language utterances with a single target speaker's voice can be easily produced using such framework and switching language-dependent prosodic models [3]. However, speaker adaptation of the separate modeling scheme has not yet been investigated.

This paper proposes an adaptation method for the regression-based prosodic models and evaluates its effectiveness in the framework of the HMM-based speech synthesis based on subjective experiments. In our prosody generation scheme, morae (Japanese syllables) or phonemes are first clustered, and then initial speaker-independent prosodic models are trained for each class using multiple speakers' utterances. The proposed speaker adaptation is conducted by adjusting bias terms for each class model so as to minimize estimation errors for adaptation data. Although the proposed method is evaluated in Japanese speech synthesis, it should be applicable to many other languages.

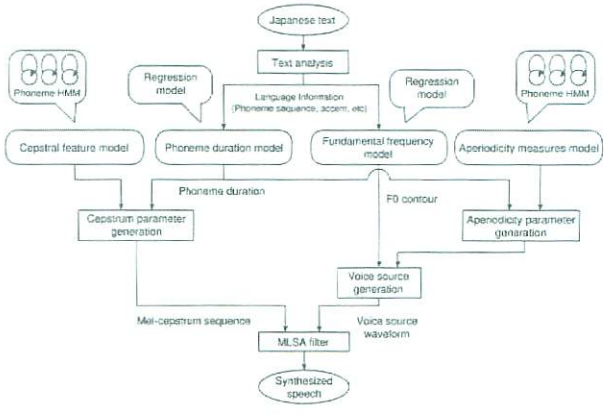

Fig. 1 HMM-based text-to-speech system in which prosodic and cepstral features are separately modeled.

The rest of the paper is organized as follows. In Section II, the prosody (F_0 contour and phoneme duration) control method using dummy-variable regression model is described. Section III proposes a prosodic model adaptation method based on bias term adjustment. Experimental conditions and results are reported in Sections IV and V respectively, and Section VI concludes this paper.

II. PROSODY GENERATION USING THE DUMMY-VARIABLE REGRESSION MODEL

In order to achieve high-quality prosody control in Japanese speech synthesis, several F_0 contour control methods [4,5,6] and duration control methods [5,7] using a dummy-variable regression model have already been proposed.

Proceedings of the Second APSIPA Annual Summit and Conference, pages 82–85,
Biopolis, Singapore, 14-17 December 2010.

A. *Dummy-variable regression model*

The dummy-variable regression model formulates the relationship between categorical and numerical values as follows:

$$\hat{y}_i = b + \sum_f \sum_c x_{fc}\delta_{fc}(i) \quad (i = 1, \cdots, N) \tag{1}$$

where $\hat{y}_i$ is a predicted value of the ith sample, b is a bias term derived as a mean value over all samples, and N is the total number of samples. $\delta_{fc}(i)$ is a characteristic function:

$$\delta_{fc}(i) = \begin{cases} 1 & : \text{ if the } i\text{th samples belong to} \\ & \quad \text{the category } c \text{ of the factor } f \\ 0 & : \text{ otherwise} \end{cases} \tag{2}$$

x_{fc} is a score of the factor f for the category c, which can be computed by minimizing the summation of squared errors $E = \sum_i (\hat{y}_i - y_i)^2$ using conventional linear regression analysis.

B. *F_0 contour generation*

In our F_0 contour generation scheme, an F_0 value at the center of each mora is predicted and an F_0 contour for a sentence is generated by linearly interpolating the moraic values [3]. The moraic F_0 values are transformed into log scaled values p by the following equation before the prediction:

$$p = 12\log_2(F_0/55) \tag{3}$$

We suppose that a target mora is included as the n-th mora in an accentual phrase W. The regression model is built separately for 5 classes of n, {1, 2, 3, 4, and larger}. The factors of the regression model are shown in Table I. In the table, () indicates the number of categories and P is a speech paragraph that W belongs to.

C. *Phoneme duration control*

Phonemes are first clustered into 13 classes as shown in Table II. In order to train the duration models, phoneme boundaries in the training utterances are estimated by the Viterbi alignment technique, using triphone-HMMs built for each speaker. Since characteristics of current, preceding/succeeding and second preceding/succeeding phonemes have been found to be effective in phoneme duration estimation in our previous experiment [8], these factors are used for phoneme duration modeling.

III. PROPOSED ADAPTATION METHOD

The regression-based prosodic models are formulated using bias terms and category scores, which control the naturalness and individuality of the synthesized voice. A regression model of F_0 contour has 181.4 category scores in average, and that of phoneme duration has 80 category scores in average. Since the F_0 contour models and the phoneme duration models are built separately for 5 classes and 13 classes respectively, the total number of bias terms is 18 and that of category scores is 1,947. A large number of free parameters cause degradation of adaptation performance when the size of adaptation data for each target speaker is limited, e.g., several utterances. Therefore, according to our preliminary experiments, we have decided to control only the bias terms and keep constant the category scores in our method to keep the robustness of adaptation[1].

A new bias term b' is estimated for each class so as to minimize the sum of the squared estimation errors for adaptation data as follows:

TABLE I
FACTORS FOR GENERATING THE F_0 CONTOUR USING THE REGRESSION MODEL.

ID	Factors (number of categories)
1	Number of moras in W (8)
2/3	Number of moras before/after W within P (9)
4	Accent type of prosodic unit W (7)
5/6	Accent type of prosodic unit before/after W (7)
7	Number of prosodic units with higher accent type than 1 before W within P (4)
8/9	Pitch connection pattern at the boundary between W and before/after W (4)
10/11	Pause length before/after W (9)
12	Tone pattern (5～10 : depending on n)
13	Current phoneme (8) (A phoneme existing at the center of each mora is either /a/, /i/, /u/, /e/, /o/, /N/, /Q/ or /-/.)
14/15	Kind of preceding/succeeding phoneme (13)
16	Position of mora M within W (if $n \geq 5$)
17	Position of accent type (if $n \geq 5$) (6)
18/19	Pause length at phrase boundary before/after W (9)
20/21	Number of moras before/after W within P (9)
22/23	Pitch connection pattern at the boundary between second preceding/succeeding W and preceding/succeeding W (5)
24/25	Pitch connection pattern at the boundary between third preceding/succeeding W and second preceding/succeeding W (5)
26/27	Kind of second preceding/succeeding phoneme (6)

TABLE II
LIST OF PHONEME CLASSES.

Phoneme classes	phoneme
1. Vowel	/a/ , /i/ , /u/ , /e/ , /o/
2. Syllabic nasal	/N/
3. Choked sound	/Q/
4. Long vowel	/-/
5. Voiced stop	/b/ , /d/ , /g/
6. Unvoiced stop	/p/ , /t/ , /k/
7. Voiced fricative	/z/ , /j/
8. Affricate	/ch/ , /ts/
9. Unvoiced fricative	/f/ , /h/ , /s/ , /sh/
10. Nasal consonant	/m/ , /n/
11. Liquid	/r/
12. Semi vowel	/w/ , /y/
13. Palatalized consonant	/by/ , /dy/ , /gy/ , /py/ , /ky/ , /hy/ , /ry/ , /my/ , /ny/

[1] We tried to expand our adaptation method to deal with not only bias values but also category scores. However, the preliminary experiments showed that the synthesized speech by this method was significantly deteriorated when only a small amount of adaptation data was used.

$$\frac{\partial E}{\partial b'} = \frac{\partial}{\partial b'}\sum_i (\hat{y}_i - y_i)^2 = 0$$
$$\Rightarrow \quad b' = \frac{1}{N'}\sum_i \left(y_i - \sum_f \sum_c x_{fc}\delta_{fc}(i) \right) \tag{4}$$

where y_i' and $\hat{y}_i'$ are correct and predicted values of the ith sample of adaptation data, respectively. N' is the total number of adaptation samples. The estimated bias value b' is not equal to the simple mean value directly calculated from the adaptation samples, because the category scores are considered in the estimation process. Since the bias adjusting operation is applied to each class model constructed considering the difference of mora positions/kind of phonemes, some prosodic individualities caused by the speaker's dialect or habit can be represented in the adapted model; e.g., a habit such as frequently raising F_0 at phrase final syllables.

IV. Experimental Conditions

In order to evaluate effectiveness of the proposed adaptation method, naturalness and individuality of the synthesized voices were measured by subjective tests.

A. Speech database and analysis condition

ATR Japanese speech database consisting of 503 sentences (Sets A to I, each including 50 sentences, and Set J, including 53 sentences) uttered by four male speakers (MHT, MYI, MTK and MMY) and four female speakers (FKS, FKN, FTK and FYM) was used for the experiment (503x8=4,024 sentences in total). They were separately used for training/adaptation and testing. After cepstral and F_0 features were extracted by the STRAIGHT method [9] under the condition of 16-ms window length and 1-ms shifting, the extracted features were smoothed and re-sampled at 5 ms periods. The cepstral triphone-HMM for extracting phoneme durations was built as a speaker-independent model trained using the Set A-I by all the eight speakers (450x8=3,600 sentences).

B. Flow of experiment

First, speaker-independent cepstral, F_0 contour generation and phoneme duration generation models were trained using the Set A-I (450 sentences) and Set J01-J20 (20 sentences) by seven speakers except for the target speaker. Then, 5 or 20 sentences by the target speaker were randomly selected from the Set A-I for adaptation. Speaker-adapted prosodic models were generated by the proposed adaptation method, and speaker-adapted cepstral model was generated by the SMAPLR method [10]. For the comparison purpose, speaker-trained prosodic and cepstral models were built by 100, 200, 400 or 470 sentences selected randomly from the Set A-I and J01-20 by the target speaker.

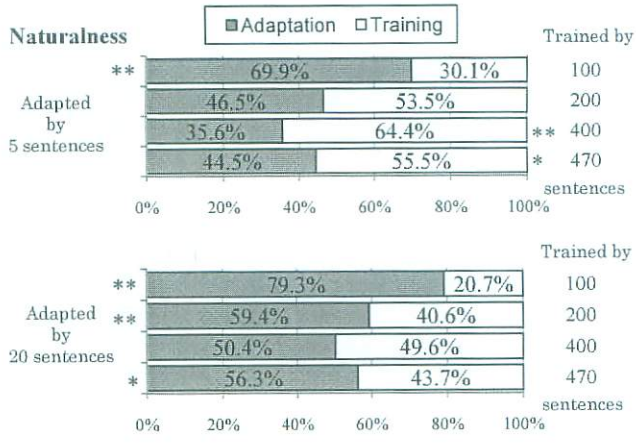

Fig.2 Preference scores for naturalness of synthesized speech produced by speaker-adapted and speaker-trained models. "*" and "**" indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

Sentences in the Set J21-J53 were converted into synthesized speech using the speaker-adapted model and the speaker-trained model, and their naturalness and individuality were evaluated by subjective tests. In the naturalness test, subjects listened to a pair of utterances with the same sentence in random order: one was synthesized by the speaker-adapted model and the other was synthesized by the speaker-trained model. The subjects chose subjectively more natural one from the two. In the individuality test, subjects first listened to a reference speech synthesized by using originally extracted target speaker's features, and chose more similar speech to the reference speech from the two synthesized speeches. 16 subjects were used in these experiments.

V. Experimental Results

A. Naturalness

Naturalness test results using 5 sentences or 20 sentences for adaptation are shown in Fig. 2. Naturalness of synthesized speech using models adapted by 5 sentences was better than that of synthesized speech using models trained by 100 sentences, almost equivalent to models trained by 200 sentences, and worse than models trained by 400 or 470 sentences. Naturalness of synthesized speech using models adapted by 20 sentences was better than that of synthesized speech using models trained by 100, 200 or 470 sentences, and almost equivalent to models trained by 400 sentences.

B. Individuality

Individuality test results are shown in Fig. 3, which are almost equivalent to that for naturalness. Individuality of synthesized speech using models adapted by five sentences was better than that of synthesized speech using models trained by 100 sentences, almost equivalent to models trained by 200 sentences, and worse than models trained by 400 or

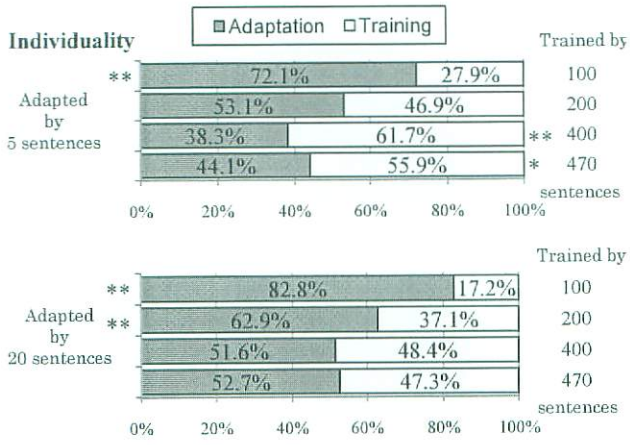

Fig.3 Preference scores for individuality of synthesized speech produced by speaker-adapted and speaker-trained models. "*" and "**" indicate that differences are statistically significant at 5% and 1% significance levels, respectively.

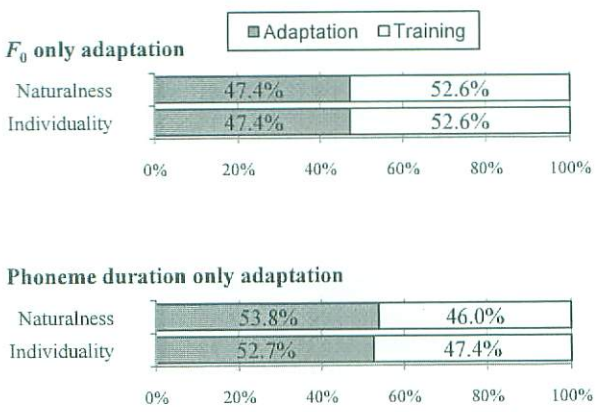

Fig. 4 Preference scores for naturalness and individuality of synthesized speech produced by a speaker-adapted model (using 5 sentences) and a speaker-trained model (using 450 or 470 sentences), when only F_0 contour or phoneme duration is adapted and other features are trained for each speaker.

470 sentences. Individuality of synthesized speech using models adapted by 20 sentences was better than that of synthesized speech using models trained by 100 and 200 sentences, and almost equivalent to models trained by 400 or 470 sentences.

C. *Prosodic feature adaptation effect*

Supplementary experiments were conducted, in which only an F_0 contour or a phoneme duration model was adapted using 5 sentences and other models were trained for each speaker. Adaptation results were compared with training results using 450 or 470 sentences. Preference scores are shown in Fig. 4. For both naturalness and individuality, the model adapted by 5 sentences was almost equivalent to the speaker-trained model. Considering the results comparing the quality of utterances by 5-sentences-adaptation and 470-sentences-training shown in Figs. 2 and 3 where all the features were updated, it can be concluded that 5 sentences are enough for the prosodic features themselves in terms of the subjective tests, although 20 sentences are needed to include cepstral HMM adaptation.

In addition to the subjective tests, we also performed objective tests for the prosodic models where their prediction is compared with true value. The results showed that the prediction error by the F_0 model that was adapted with 5 sentences was roughly equivalent to the one that was trained with 450 sentences. On the other hand, the duration model had significantly larger error when adapted with 5 sentences than the model trained with 450 sentences. Duration model adapted with 20 sentences had roughly an equivalent error to the one trained with 100 sentences.

VI. Conclusions

In order to synthesize speech with high naturalness and individuality using only a small amount of adaptation utterances from a target speaker, this paper has proposed F_0 contour and phoneme duration generation methods based on dummy-variable regression models in which bias terms for 5 and 13 classes respectively are adjusted. Subjective evaluation test results confirm that naturalness and individuality of the synthesized speech using models adapted by 5 sentences are equivalent to that using models trained by 450 or 470 sentences. When not only prosodic features but also cepstral HMM are adapted, adaptation by 20 sentences is better than training by 100 and 200 sentences, and is almost equivalent to training by 400 or 470 sentences. These results indicate that the proposed method for F_0 contour generation and the phoneme duration modeling can effectively adapt models in terms of naturalness and individuality.

In this paper, only read speech was used for evaluation. As future research, we are planning to use widely varied speech, such as emotional and spontaneous speech, and confirm the effectiveness of the proposed method. Investigating other formulations for the prosodic modeling, such as utilizing a log-scaled duration model to avoid negative values, is also a future work.

References

[1] T. Masuko, et al., *IEICE Trans. D-II* vol. J83-D-II, no.7, pp.1600-1609 (2000-7). (in Japanese)
[2] C.Hayashi, Ann. Inst. Statist. Math., vol.3, no.2, pp.69-98, 1952.
[3] J. Latorre, et al., *Proc. EUROSPEECH2005*, Lisbon, Portugal, pp.1477-1480 (2005-9).
[4] T.Sakayori, et al., *Proc. ASJ Autumn Meeting 86*, vol. 1, pp.245-246, (1986-10). (in Japanese)
[5] M. Abe, et al., *Proc ICASSP 93*, vol. 2, pp. 53-56 (1992-3).
[6] N. Kaiki, et al., *IEICE Trans. D-II*, vol. J83-D-II, no.9, pp.245-246 (1986-10). (in Japanese)
[7] N. Kaiki, et al., *Proc. ICSLP 90*, vol. 1, pp.17-20 (1990-11).
[8] K. Iwano et al, *Proc. TTS2002*, Santa Monica, U.S.A (2002-9).
[9] H. Kawahara et al., Speech Communication, vol.27, pp.187-207, 1999.
[10] O. Shiohan, et al., Computer Speech & Language, vol.16, no.3, pp.5-24, 2002.