HMM-BASED EXPRESSIVE SPEECH SYNTHESIS BASED ON PHRASE-LEVEL F0 CONTEXT LABELING

Yu Maeno[1]*, Takashi Nose*[1]*, Takao Kobayashi*[1]*, Tomoki Koriyama*[1]*,*
Yusuke Ijima[2]*, Hideharu Nakajima*[2]*, Hideyuki Mizuno*[2]*, Osamu Yoshioka*[2]

[1]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
[2]NTT Media Intelligence Laboratories, NTT Corporation
maeno.y.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

ABSTRACT

This paper proposes a technique for adding more prosodic variations to the synthetic speech in HMM-based expressive speech synthesis. We create novel phrase-level F0 context labels from the residual information of F0 features between original and synthetic speech for the training data. Specifically, we classify the difference of average log F0 values between the original and synthetic speech into three classes which have perceptual meanings, i.e., high, neutral, and low of relative pitch at the phrase level. We evaluate both ideal and practical cases using appealing and fairy tale speech recorded under a realistic condition. In the ideal case, we examine the potential of our technique to modify the F0 patterns under a condition where the original F0 contours of test sentences are known. In the practical case, we show how the users intuitively modify the pitch by changing the initial F0 context labels obtained from the input text.

Index Terms— HMM-based expressive speech synthesis, prosodic context, unsupervised labeling, audiobook, prosody control

1. INTRODUCTION

HMM-based speech synthesis has become one of the promising approaches to synthesizing natural-sounding expressive speech in TTS applications [1]. It has been shown that the typical emotional expressions, speaking styles, and emphasis expressions are well modeled by using style-dependent modeling [2] or by introducing the style/emphasis information into the context labels (e.g., [2, 3]). However, this approach can be used only in the case where the para-linguistic information of each utterance or segment is known.

Recently, the demands are increasing for synthesizing more expressive speech from speech corpora such as audiobook data with diverse and multiple style expressions [4]. For this purpose, it tends to be expensive to adopt the traditional technique straightforwardly where manual labeling of style information is required. Moreover, there is another problem that annotation results of perceived style expressions much depend on annotators, and inter-rater agreement is not always satisfactory. To avoid these problems, unsupervised data-driven style clustering technique have been proposed [4, 5]. In [5], Sźekely et al. applied self-organizing feature maps to voice quality parameters for clustering the audiobook speech data in the unit-selection synthesis. In [4], hierarchical k-means clustering was employed as the clustering method in the HMM-based synthesis. However, a crucial problem is that there are not always reliable perceptual or para-linguistic meanings in the resultant clusters, and hence it would be not always possible for users to choose an appropriate cluster and output the desired expressive speech in the speech synthesis process.

In this paper, we propose an alternative approach to enriching the prosodic variations of synthetic speech for HMM-based expressive speech synthesis. We expand the idea of our previous study [6] where we described a labeling technique of emphasis expressions. In the proposed technique, we define a novel three-class pitch context called *phrase-level F0 context* (or simply F0 context) of low, normal, and high by classifying the difference of log F0 values between the original and synthetic speech samples at a phrase level. In contrast to the other data-driven style clustering approaches [4, 5], the proposed technique can provide context labels having clear perceptual meanings and hence users can easily modify the pitch in a practical application.

2. EMPHASIS/NON-EMPHASIS CONTEXT LABELING AND ITS PROBLEMS

In our previous study [6], we focused on the unsupervised labeling of emphasis/non-emphasis accent phrases using F0 features for Japanese appealing-style speech, which is a two-class classification of prosodic patterns. Specifically, we temporarily generate F0 patterns from HMMs trained using context labels without the emphasis context as shown in Fig. 1. The accent phrase boundaries are shown with vertical dashed lines. From the figure, it is seen that in the first accent phrase the F0 prominence in the original speech is not well reproduced and the average F0 value of the synthetic speech is clearly lower than that of the original one. We calculate the difference d of the average log F0 values between original and synthetic speech for each accent phrase and classify the phrase as emphasized when d is larger than a pre-determined threshold.

Although the objective and subjective evaluation results in [6] showed the effectiveness of the above labeling technique, this approach has two problems. First, an appropriate classification threshold must be determined in advance, however, it is difficult to automatically obtain such a threshold and a manually chosen fixed threshold was used in [6]. To apply this approach to a various types of stylized speech of arbitrary speakers, we should determine an optimal threshold for each training data set because the appropriate threshold could vary depending on speakers and styles. The second problem is that the technique can be used only in the case where the average F0 values of emphasized phrases are higher than those of non-emphasized phrases. For example, when a user synthesizes speech and feels that a certain phrase should have lower pitch, the technique cannot meet this request.

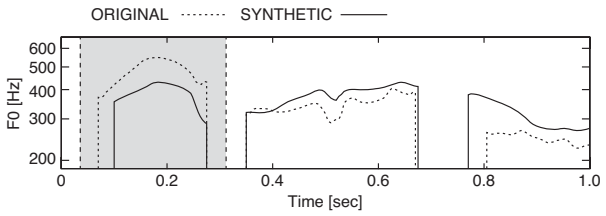

Fig. 1. Example of F0 patterns of original (ORIGINAL) and synthetic speech (SYNTHETIC) without the emphasis context for appealing style speech.

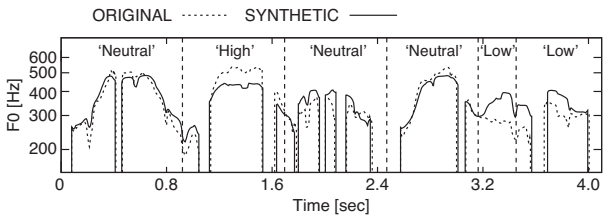

Fig. 2. Example of F0 context labels obtained in model training for appealing speech.

3. THREE-CLASS F0 CONTEXT LABELING WITH THRESHOLD OPTIMIZATION

As described in Section 2, we need more appropriate labeling technique to enrich the prosodic variations of synthetic speech with arbitrary speakers and styles. Therefore, we modify the conventional emphasis/non-emphasis labeling approach to three-class F0 context labeling which allows users to modify the relative pitch of synthetic speech, i.e., higher or lower pitch at the phrase-level.

3.1. Context labeling using phrase-level prosody difference

For creating the phrase-level F0 context labels, we use the difference of the average log F0 values of each accent phrase of original and synthetic speech [6]. The F0 context labeling process is summarized as follows:

1. Train context-dependent HMMs using conventional labels obtained from linguistic information only.
2. Generate F0 sequences for the training sentences using the trained HMMs.
3. Calculate average log F0 values f_o and f_s of original and synthetic speech for each accent phrase.
4. Calculate the average log F0 difference $d = f_o - f_s$.
5. Classify the difference d into three classes: (1) $d < -\alpha$ (low), (2) $-\alpha \leq d < \alpha$ (neutral), and (3) $d \geq \alpha$ (high), where positive value α is a classification threshold.
6. The class index is used as the F0 context label.

The difference d can have both negative and positive values, which is used to determine whether the average log F0 of synthetic speech of the accent phrase is lower or higher than that of the original speech. Figure 2 shows an example of obtained F0 context labels for a training utterance of appealing speech with original and generated F0 contours. From the figure, the F0 contour generated using the conventional technique is flattened and differs greatly from that of the original speech in the second, fifth, and sixth accent phrases. The F0 context describes these differences and compensates the conventional context labels that include only linguistic information. It should be noted that using even numbers of classes, such as two, has a problem in the labeling because there is no class index corresponding to the neutral class. In this study, although we define the F0 context for each accent phrase that might be particular to Japanese speech, we may use other unit such as word- or syllable-level unit. However, using a smaller unit such as phone-level unit would be not appropriate for practical applications since there are too many F0 context labels to be set even in one utterance.

3.2. Optimization of classification threshold

In the proposed technique, we determine the appropriate classification threshold α using the training data. We choose the optimal α that minimizes an objective measure using a simple grid search. As the objective measure, we employ root mean square (RMS) error of log F0 between original and synthetic speech. The process for optimizing the threshold is summarized as follows:

1. Choose a value of α between pre-determined lower and upper bounds, α_s and α_e.
2. Classify the average log F0 difference d of the accent phrase into one of three classes using the specified α, and create F0 context labels for all training sentences.
3. Train new HMMs using the labels including the obtained F0 context, and then generate F0 sequences for all training sentences using the F0 context labels in addition to the ordinary context labels.
4. Calculate the RMS error E_α of log F0s for the threshold α between original and newly synthesized speech.
5. Change the value of α ($\alpha_s \leq \alpha \leq \alpha_e$) and repeat the steps from 1 to 4.
6. Finally, obtain the optimal threshold α^* that minimizes E_α:

$$\alpha^* = \arg\min_{\alpha} E_\alpha. \quad (1)$$

3.3. Use of F0 context labels in synthesis phase

A problem in the proposed technique is that we cannot obtain the proposed F0 context labels from the input text automatically in the speech synthesis process since the labels are determined from the original expressive speech utterance. This type of problem of the data-driven expressive speech classification in speech synthesis studies often remains unsolved and is supposed to be discussed in the future work [4, 5]. In this paper, we provide a practical usage of the proposed technique under a realistic condition where unknown context information for the input sentences is not used. A typical application is to create synthetic speech samples of voice actors/actresses with higher prosodic variebility for audiobook and movie contents. In the condition, users first synthesize speech for the target sentence using F0 context labels whose values are all set to neutral. The synthetic speech obtained by this process has sometimes poor prosodic variability compared to the natural speech. Next, the users listen to the speech sample and change the F0 context of a certain accent phrase to high/low if they want to modify the pitch of the sentence. Then they synthesize speech again using the modified labels and check the resultant synthetic speech. By repeating this procedure, users can obtain the better synthetic speech in terms of F0 variations.

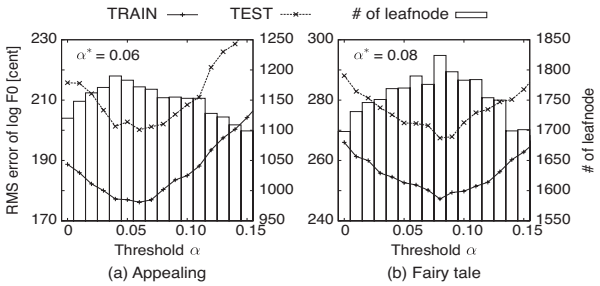

Fig. 3. RMS errors of log F0 and the numbers of leaf nodes with different classification thresholds for the first data set of cross-validation.

4. EXPERIMENTS

4.1. Experimental conditions

In the following experiments, we used two types of Japanese speech data, appealing and fairy tale speech uttered by a female and a male professional narrators, respectively. The appealing speech was taken from the database described in [7], where a female salesclerk speaks to her customers to push some products through mass media commercials. The amounts of speech utterances of appealing and fairy tale speech are approximately 33 and 52 minutes, respectively. Speech signals were sampled at a rate of 16kHz and the frame shift was 5 ms. We used STRAIGHT analysis [8] for speech feature extraction, and extracted spectral envelope, F0, and aperiodicity features. The spectral envelope was then converted to mel-cepstral coefficients using a recursion formula. The aperiodicity feature was also converted to average values for five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. As a result, the feature vector consisted of 39 mel-cepstral coefficients including the zeroth coefficient, log F0, five-band aperiodicity values, and their delta and delta-delta coefficients. The total dimensionality was 138. We used five-state left-to-right hidden semi-Markov model with no skip topology. The output distribution in each state was modeled with a single Gaussian density function, and covariance matrices of these models were assumed to be diagonal. In the context clustering for parameter tying, decision trees were automatically constructed based on the minimum description length (MDL) criterion [9]. For the respective speakers, we conducted four-fold cross-validation. Prosodic variation could affect not only the current phrase but also the adjacent phrases, hence we take into account the preceding and succeeding accent phrases as well as the current one for the proposed F0 context.

4.2. Optimal thresholds for respective data sets

First, we show how the optimal threshold for the F0 context was determined in the model training. In this experiment, α_s were fixed to zero, and α_e was set on the basis of the maximum value of d for all accent phrases of the training data. In the first iteration, we set $\alpha = \alpha_s$, and in the n-th iteration, we set α as follows:

$$\alpha = \alpha_s + (n-1) \cdot \Delta\alpha \quad (\alpha \leq \alpha_e) \tag{2}$$

where $\Delta\alpha$ is an increment in each iteration. Here, we set α_e to 0.3 and $\Delta\alpha$ to 0.01. Figure 3 shows the RMS errors of log F0 and the average numbers of total leaf nodes of F0 trees with different values of α. In the figure, the RMS log F0 errors calculated from the test data (TEST) as well as the training data (TRAIN) are shown for $0 \leq \alpha \leq 0.15$. From the figure, we see that the optimal thresholds were also suitable for test data. In addition, the numbers of leaf nodes became largest around the optimal thresholds, which indicates that the prosodic variations are efficiently represented by the proposed F0 context. The optimal thresholds for respective data sets of two speakers in four-fold cross-validation are shown in Table 1. We see that optimal threshold varies depending on the type of speech by comparing the results of appealing and fairy tale speech. In addition, the choice of the data set seems to affect the result in the fairy tale speech.

Table 1. Optimal threshold α^* for each data sets in four-fold cross-validation, where 0.06 = 104 cent, 0.08 = 138 cent, and 0.10 = 173 cent.

	Data set			
Type	1	2	3	4
Appealing	0.06	0.06	0.06	0.06
Fairy tale	0.08	0.10	0.10	0.08

Table 2. RMS errors of log F0 [cent] between original and synthetic speech for test data.

Type	BASELINE	CONVENTIONAL	PROPOSED
Appealing	253.9	222.5	**201.2**
Fairy tale	358.9	295.4	**272.8**

4.3. Evaluation in an ideal condition

Before evaluating the proposed technique in a practical situation, we investigated the potential capability of the proposed technique in an ideal condition where we created the F0 context labels using the same way as the labeling for the training data. Specifically, we temporarily generated the speech parameters for the test sentences, and calculated the differences of average values of log F0 for each accent phrase between synthetic and original speech samples. Then, we created the F0 contexts by classifying the differences into three classes using the optimal threshold α^* obtained from the training data.

We objectively evaluated the effect of the F0 context using the RMS error of log F0 between original and synthetic speech. We compared the following three techniques.

BASELINE: Using labels without phrase-level F0 context.

CONVENTIONAL: Using labels with F0 context obtained by the conventional two class classification [6] with fixed threshold (α = 0.058 = 100 cent).

PROPOSED: Using labels with F0 context obtained by the proposed three-class classification with the optimal thresholds for respective data sets and speakers.

Table 2 shows the results. From the table, we see that the RMS errors of log F0 were significantly decreased by using the proposed technique compared to the baseline and conventional ones in both types

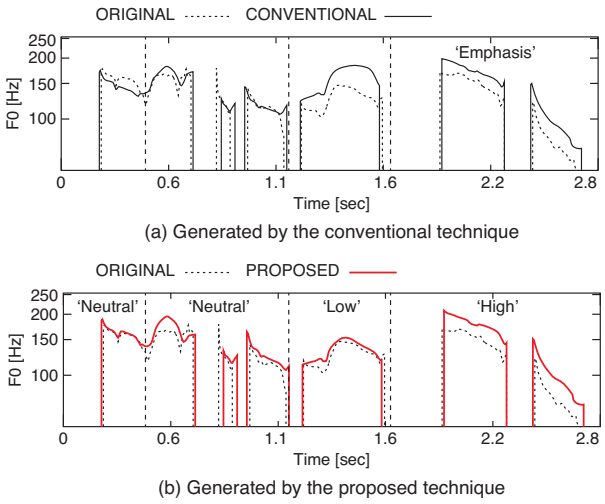

Fig. 4. Example of F0 contours of fairy tale speech generated with and without ideal F0 context labels.

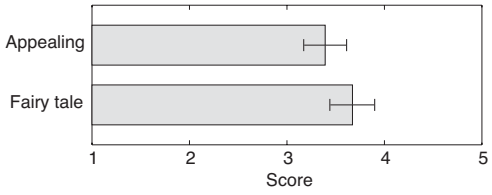

Fig. 5. Results of satisfaction rating on modifying pitch of accent phrases.

of speech. Figure 4 shows an example of F0 contours of original and synthetic speech by the conventional and the proposed techniques for fairy tale speech. From the figure, it is shown that the F0 reproducibility can be improved by appropriately setting the F0 context of high/low in the synthesis labels of each accent phrase.

4.4. Evaluation in a practical situation

We conducted a listening test under a more realistic condition to examine whether users can actually obtain desired synthetic speech with intended pitch variations by using the procedure described in Section 3.3. The number of participants was nine. First, participants listened to a reference sample (NEUTRAL) synthesized using labels where all F0 context labels were set to neutral and chose an arbitrary accent phrase whose pitch they want to be higher than the current one. Then, they listened to the reference sample and a re-synthesized sample (MODIFIED) generated using the modified F0 context labels they newly specified, and rated the satisfaction level for the pitch modification using a five-point scale: "1" for bad, "2" for poor, "3" for fair, "4" for good, and "5" for excellent.

Figure 5 shows the average scores of the listening test with confidence intervals of 95%. The average scores of both types of speech are between fair and good, which indicates the effectiveness of the

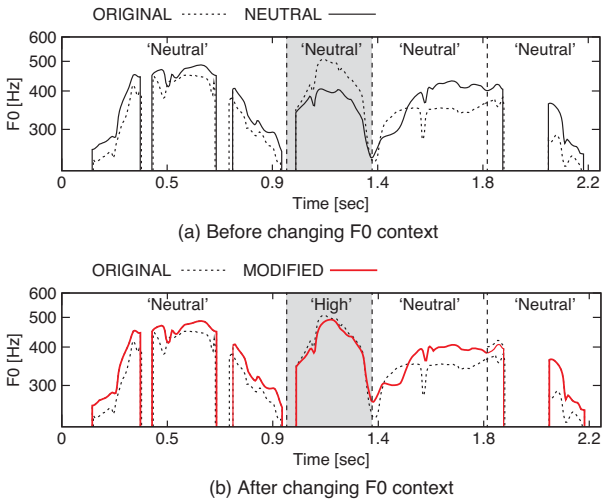

Fig. 6. Example of F0 contours before and after changing F0 context in appealing speech. The chosen accent phrase is shown as the shaded region.

proposed technique in a practical use. Figure 6 shows an example of F0 contours of the original speech as ORIGINAL and speech before and after the F0 modification as NEUTRAL and MODIFIED, respectively. In the label modification process, first the participant listened to the synthetic speech whose F0 pattern is shown as NEUTRAL in Fig. 6 (a). The participant felt a lack of prominence in second accent phrase, hence it was chosen to be modified to higher pitch. In Fig. 6 (b), we can see the resultant F0 contour became closer to that of the original speech. The rating score for this modification was four.

5. CONCLUSIONS

In this paper, we have proposed an unsupervised labeling of phrase-level F0 context that can be used for enhancing the HMM-based expressive speech synthesis. The additional three-class F0 context is determined at the phrase level by the classification of the difference of average log F0 values of original and synthetic speech for the training data. From the objective and subjective evaluation results, we confirmed that we can significantly improve the reproducibility of the expressivity of the target speech by appropriately setting the F0 context in the speech synthesis of input sentences. The first experiment was done under an ideal condition where the original F0 patterns were assumed to be known, the results showed the potential ability of the proposed technique for modifying the prosodic characteristics of the synthetic expressive speech. The result of the second experiment conducted under a more practical situation showed that the users can intuitively modify the prosodic characteristics of a target accent phrase by changing the F0 context from neutral to high. The future work will include an application of proposed labeling technique to other prosodic features such as power and duration.

6. REFERENCES

[1] T. Nose and T. Kobayashi, "Recent development of HMM-based expressive speech synthesis and its applications," in *Proc. APSIPA ASC 2011*, 2011, http://www.apsipa.org/proceedings_2011/pdf/APSIPA189.pdf.

[2] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, 2005.

[3] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden Markov models," in *Proc. Oriental COCOSDA*, 2009, pp. 76–81.

[4] F. Eyben, S. Buchholz, and N. Braunschweiler, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Proc. ICASSP 2012*, 2012, pp. 4009–4012.

[5] E. Sźekely, J. Cabral, P. Cahill, and Carson-Berndsen J., "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proc. INTERSPEECH 2011*, 2011, pp. 2409–2412.

[6] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based emphatic speech synthesis using unsupervised context labeling," in *Proc. INTERSPEECH 2011*, 2011, pp. 1849–1852.

[7] H. Nakajima, N. Miyazaki, A. Yoshida, T. Nakamura, and H. Mizuno, "Creation and analysis of a Japanese speaking style parallel database for expressive speech synthesis," in *Proc. Oriental COCOSDA*, 2010, http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf.

[8] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[9] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, pp. 79–86, 2000.