HMM-Based Emphatic Speech Synthesis Using Unsupervised Context Labeling

Yu Maeno[1], *Takashi Nose*[1], *Takao Kobayashi*[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 Cyber Space Laboratories, NTT Corporation

maeno.y.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper describes an approach to HMM-based expressive speech synthesis which does not require any supervised labeling process for emphasis context. We use appealing-style speech whose sentences were taken from real domains. To reduce the cost for labeling speech data with an emphasis context for the model training, we propose an unsupervised labeling technique of the emphasis context based on the difference between original and generated F0 patterns of training sentences. Although the criterion for the emphasis labeling is quite simple, subjective evaluation results reveal that the unsupervised labeling is comparable to the labeling conducted carefully by a human in terms of speech naturalness and emphasis reproducibility.

Index Terms: HMM-based speech synthesis, expressive speech, emphasis expression, unsupervised labeling, F0 generation

1. Introduction

One of the goals of text-to-speech synthesis is to generate human-like expressive speech which can express various paralinguistic information such as emotion, intention, and speaking style. In this context, it is well known that prosodic features play an essential role. To reproduce the prosodic variations of expressive speech, HMM-based speech synthesis [1] is a promising approach because of its flexibility in the modeling and parameter generation. For instance, we have shown that emotional expressions and speaking styles, which we refer to as *styles*, are well modeled using an HMM-based framework [2]. In the conventional studies, we have focused mainly on reproducing global style characteristics which consistently appear in whole parts of speech samples of a target style. For such stylized speech, the conventional contextual factors work well in the modeling and synthesis as well as the reading-style speech.

On the other hand, there is another typical expression, i.e., *emphasis* expression in expressive speech. Emphasis is important to correctly communicate our intention to the others in speech communication. The emphasis has different property with emotions and speaking styles since the emphasis expressions appear partially in an utterance. Therefore, modeling and generation of such locally emphasized speech is difficult when we use the conventional context set having no emphasis information. There have been several studies to synthesize the emphatic speech [3–6] in the HMM-based speech synthesis framework. In [3], it was reported that an emphasis context worked well under a constraint that only a single word was emphasized in reading-style speech recorded using emphasis-directed script. More unconstrained corpus was also evaluated where the contrast expression in dialogue speech were focused on using automatically detected contrastive word pairs from texts [4]. A problem of these simple context-based techniques is that the emphasis can not be well modeled when the expressions are very weak. For such natural emphatic speech, the context adaptive training with factorized decision trees [6] has been shown to be effective.

In this paper, we focus on the F0 feature which has been considered to be one of the primary factors of emphasis expression [7, 8], and investigate the HMM-based emphatic speech synthesis using more expressive and unconstrained speech than previous studies [3–5]. We use appealing-style speech of a Japanese female speaker, where a lot of emphasis expressions are included in respective utterances. We focus on F0 prominences appearing in the emphasis expressions since Japanese is a pitch accent language and the F0 prominences are the most typical characteristics among prosodic variations. We also propose an unsupervised context labeling technique of emphatic speech data by utilizing the property of F0 patterns generated without an emphasis context. We conduct objective and subjective evaluation experiments to examine the effectiveness of the proposed unsupervised labeling.

2. Unsupervised emphasis context labeling

2.1. F0 prominence as an emphatic expression in expressive speech

In this study, we used a female speaker's appealing-style speech taken from the Japanese speech database used in [9]. The database also includes reading-style speech of the same sentences uttered by the same speaker as those of the appealing-style. In the recording of appealing-style speech, sentences were taken from real domains. A female salesclerk speaks to her customers to push some products through mass media commercials. It is noted that the phrases to be emphasized was not directed through the recording. The total number of sentences is 248, and there are 2186 accent phrases.

As a reference, we manually labeled the appealing-style speech samples with an emphasis symbol for each accent phrase. We use accent phrase as a labeling unit of the emphasis context since Japanese is pitch-accent language and the accent phrase is a basic unit to express the pitch accent. The suitable unit would depend on the target language, e.g., a word unit is used for English speech in [4, 5]. The labeling was conducted by one of authors of this paper. Figure 1 shows an example of F0 patterns of reading- and appealing-style speech extracted from speech samples taken from the database. In the figure, the emphasized accent phrase is shown as a colored region. We can see that there is a clear difference of F0 patterns between two styles. As global characteristics, the appealing-style gives

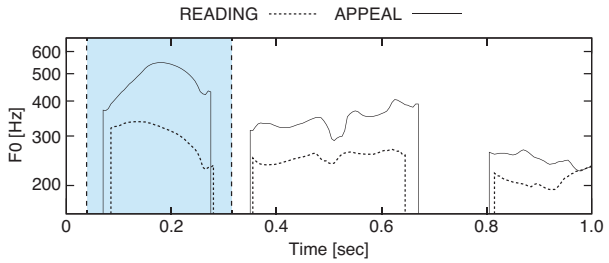

Figure 1: Local variation of the F0 pattern in an emphasized phrase. The emphasized accent phrase is shown as a colored region.

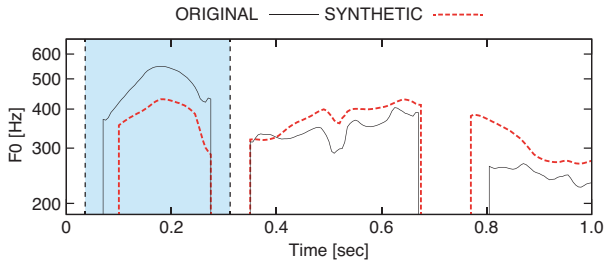

Figure 2: Difference between original and generated F0 patterns without an emphasis context. The emphasized accent phrase is shown as a colored region.

higher F0 values than the reading-style. Moreover, the F0 difference in the emphasized phrase between two styles becomes larger than that of the unemphasized phrases. To confirm that such an F0 prominence in emphasized phrases can be seen in whole speech utterances, we calculated the differences of mean log F0 values between emphasized and unemphasized accent phrases of appealing-style speech. The results were 399 and 351 cents in emphasized and unemphasized phrases[1], and we found the two means are statistically significantly different at a 1% level.

2.2. Generated-F0-based unsupervised emphasis labeling

The local variation described in the previous section is difficult to be modeled by the conventional HMM-based speech synthesis where an emphasis context is not taken into account[2]. Figure 2 shows an example of the F0 pattern generated from the model trained using the appealing-style speech without an emphasis context. We can see that the emphasis was not well reproduced and the F0 values of the synthetic speech became consistently lower than those of the original speech in the emphasized accent phrases. Taking into account this behavior of the HMM-based speech synthesis, we attempt to automatically label the emphasized phrases with an emphasis context. The labeling process is summarized as follows:

1. Train context-dependent HMMs of appealing-style speech using conventional labels without an emphasis context.
2. Generate F0 sequences using the training sentences.
3. Calculate mean log F0 values f_o and f_s of original and synthetic speech for each accent phrase.
4. Calculate the difference $d = f_o - f_s$.
5. If the difference d is larger than a pre-determined threshold, the phrase is labeled as emphasized.

The basic idea of the proposed labeling technique is the same as [10] where the generated F0 pattern from HMMs was used for the automatic detection of an emphasis expression. However, the detection technique in [10] is based on supervised training and a sufficient amount of manually labeled speech data[3] is required for the emphasis prediction. On the other hand, our technique needs no supervised data labeled by a human since the labeling is done based on a simple threshold operation.

Table 1: Evaluated context labels.

Label	Contextual factor
NORMAL	phoneme, accent, sentence length
EMPHASIS_A	NORMAL + emphasis (automatic)
EMPHASIS_M	NORMAL + emphasis (manual)

2.3. Context label

In the following experiments, we used three types of context labels shown in Table 1. NORMAL is the conventional label set including contextual factors related to the numbers and the positions of phoneme, accent, and sentence length. EMPHASIS_A and EMPHASIS_M are label sets in which manual and automatic emphasis contexts are added to NORMAL, respectively. We used binary information of emphasis as a contextual factor. The emphasis could affect not only the current phrase but also the adjacent phrases [3], hence we take into account the preceding and succeeding accent phrases as well as the current one. The other processes of the training and synthesis are the same as those of the standard HMM-based speech synthesis [11].

3. Experiments

3.1. Experimental conditions

We performed a 5-fold cross validation test using the speech data described in Sect. 2.1. As for the test sentences, we used labels with the manually obtained emphasis context for the test speech samples. Speech signals were sampled at a rate of 16kHz and the interval of frame shift was 5 ms. We used STRAIGHT analysis [12] 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, 5-band aperiodicity values, and their delta and delta-delta coefficients. The total dimensionality was

[1] 1 octave = 1200 cents, 1 semitone = 100 cents.

[2] This is not true when the emphasis realization and conventional contextual factors are dependent on each other. When a speaker expresses the emphasis completely depending on the conventional context, e.g., he/she always emphasizes the first accent phrases, the emphasis can be reproduced without the emphasis context.

[3] They used 180 utterances for the training data.

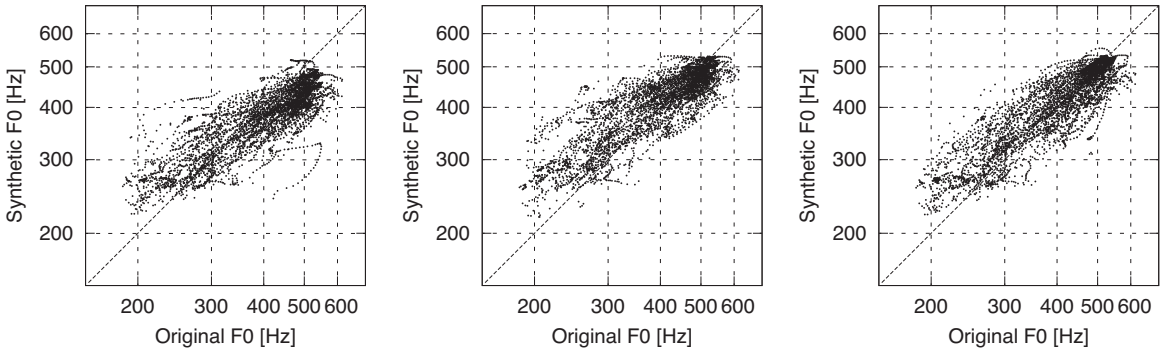

Figure 3: Distributions of generated F0 with and without an emphasis context against original F0.

Table 2: RMS errors and correlations of log F0 between original and synthetic speech for all and emphasized accent phrases.

	All		Emphasis	
Context	RMSE[cent]	Corr.	RMSE[cent]	Corr.
NORMAL	265.2	0.772	317.3	0.805
EMPHASIS_A	247.0	0.810	257.0	0.826
EMPHASIS_M	244.6	0.816	248.0	0.842

138. We used 5-state left-to-right HSMM 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, a decision tree was automatically constructed based on the minimum description length (MDL) criterion [13]. For the unsupervised labeling of emphasized accent phrases, we set the threshold described in Sect. 2.2 to 100 cents on a basis of preliminary objective and subjective experimental results.

3.2. Objective evaluations

We objectively evaluated the similarity of the generated F0 against the original one. As the objective measures, we used RMS errors and correlation coefficients of log F0 values between original and synthetic speech samples. Table 2 shows the results. In the table, the results are shown both for all and emphasized accent phrases. By adding the emphasis context, RMS errors decreased especially in the emphasis phrases. It is interesting that the correlation increased significantly not only for the emphasis phrases but also for whole phrases. Although the performance of the proposed unsupervised labeling was slightly lower than the manual labeling, the reproducibility was significantly improved compared to the case without the emphasis context.

Figure 3 shows distributions of generated F0 values of original and synthetic speech for emphasis phrases in test sentences. We chose one evaluation data set in the cross-validation, and plotted the original and generated F0 values for horizontal and vertical axes in log scale, respectively. When we used the conventional context (Fig. 3(a)), most of the generated F0 values were lower than 500Hz whereas there are many samples higher than 500Hz in the original F0. On the other hand, the F0 values higher than 500Hz was also generated by using the emphasis context (Fig. 3(b) and (c)). Figure 4 shows an example of F0 patterns generated using respective context labels. From the figure, we can see that the F0 pattern of the synthetic speech became closer to that of the original one.

3.3. Subjective evaluations

We conducted two types of subjective evaluation tests. To focus on the evaluation of F0 reproducibility, we used acoustic features extracted from the original speech except for the F0 feature. For phoneme durations, we used the correct durations included in the database. In this experiment, we randomly chose 50 test sentences from 197 sentences where at least one emphasized accent phrase is included. Then, in each test, ten sentences were randomly chosen for each participant from the 50 test sentences. The number of participants was seven.

First, we evaluated the naturalness of the synthetic speech by a MOS test. Participants rated the naturalness of speech samples on a five-point scale, i.e., 1 for bad, 2 for poor, 3 for fair, 4 for good, and 5 for excellent. The scores are shown in Fig. 5 with confidence intervals of 95%. From the results, it is found that the naturalness of the synthetic speech with an emphasis context is comparable to or slightly better than that with the conventional context. Between EMPHASIS_A and EMPHASIS_M, there was no significant difference.

Next, we evaluated whether the emphasis context improves the reproducibility of emphasis expressions appearing in the synthetic speech. As the reference samples, we used the vocoded speech of the test sentences. Participants listened to the reference and test speech samples and rated the reproducibility of emphasis expressions of the test sample by comparing it to that of the reference sample. The rating was performed using a 5-point scale, i.e., 1 for bad, 2 for poor, 3 for fair, 4 for good, and 5 for excellent. Figure 6 shows the results. It is seen that the use of the emphasis context significantly improved the reproducibility of emphasis expressions of original speech. It is also seen that the effectiveness of the automatic labeling is comparable to that of the manual labeling. A possible reason is that the criterion of the proposed emphasis labeling is consistent for whole training data.

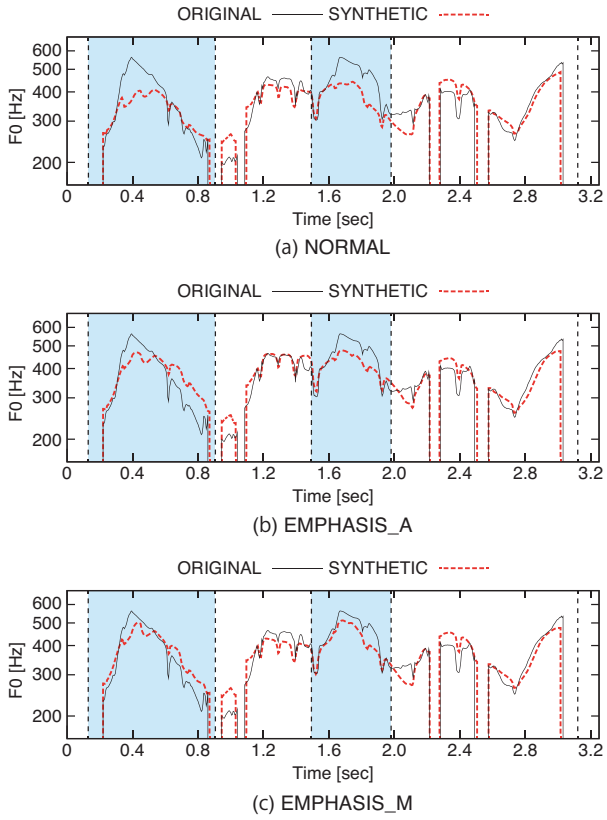

Figure 4: Example of F0 patterns generated with and without emphasis context. The emphasized accent phrases are shown as colored regions.

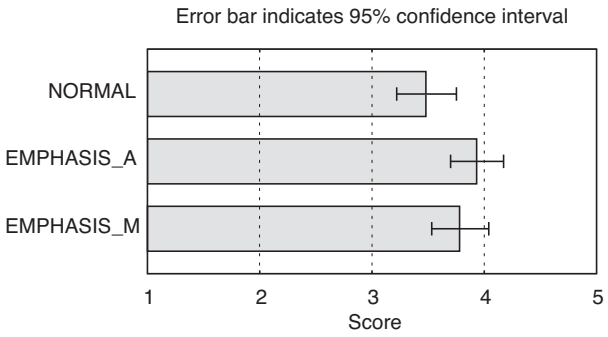

Figure 5: MOS test on the naturalness of synthetic speech.

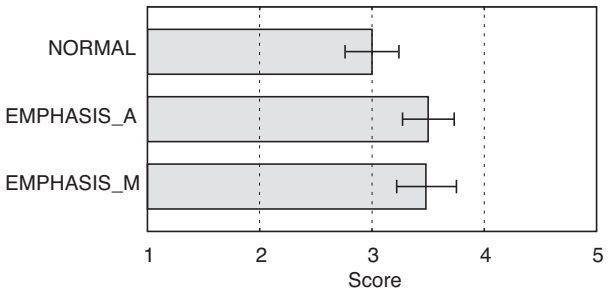

Figure 6: MOS test on the reproducibility of emphasis expressions.

4. Conclusions

In this paper, we have evaluated HMM-based emphatic speech synthesis using expressive speech including many emphasis expressions appearing as F0 prominences. We proposed unsupervised labeling of training data with an emphasis context using the property of the difference between original and generated F0 patterns. From the experimental results, we confirmed that the unsupervised labeling works quite well and is comparable to the manual labeling in terms of subjective reproducibility of emphasis. In the future work, we will investigate the other prosodic features such as power and duration of speech. It will be also important to examine the performance of the proposed labeling using other types of speech data, e.g., operators' speech and fairy tale speech included in [9].

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Sep. 1999, pp. 2347–2350.

[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] L. Badino, J. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in HMM based speech synthesis," in *Proc. INTERSPEECH 2009*, 2009, pp. 520–523.

[5] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. ICASSP 2010*, 2010, pp. 4238–4241.

[6] K. Yu, H. Zen, F. Mairesse, and S. Young, "Context adaptive training with factorized decision trees for HMM-based speech synthesis," in *Proc. INTERSPEECH 2010*, 2010, pp. 414–417.

[7] J. Brenier, D. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proc. INTERSPEECH'2005 - Eurospeech*, 2005, pp. 3297–3300.

[8] D. R. Ladd and R. Morton, "The perception of intonation emphasis: Continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[9] 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, 30, http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf.

[10] J. Xu and L. Cai, "Automatic emphasis labeling for emotional speech by measuring prosody generation error," in *Proc. ICIC 2009*, 2009, pp. 177–186.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW6*, 2007, pp. 294–299.

[12] 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, Sep. 1999.

[13] 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, Mar. 2000.