Journal of Computer and Communications, 2016, 4, 79-94
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Prosodically Rich Speech Synthesis Interface Using Limited Data of Celebrity Voice

Takashi Nose[1], Taiki Kamei[2]

[1]Department of Communication Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan
[2]Department of Applied Information Sciences, Graduate School of Information Sciences, Tohoku University, Sendai, Japan
Email: tnose@m.tohoku.ac.jp

How to cite this paper: Nose, T. and Kamei, T. (2016) Prosodically Rich Speech Synthesis Interface Using Limited Data of Celebrity Voice. *Journal of Computer and Communications*, **4**, 79-94.
http://dx.doi.org/10.4236/jcc.2016.416006

Received: October 16, 2016
Accepted: December 26, 2016
Published: December 30, 2016

Abstract

To enhance the communication between human and robots at home in the future, speech synthesis interfaces are indispensable that can generate expressive speech. In addition, synthesizing celebrity voice is commercially important. For these issues, this paper proposes techniques for synthesizing natural-sounding speech that has a rich prosodic personality using a limited amount of data in a text-to-speech (TTS) system. As a target speaker, we chose a well-known prime minister of Japan, Shinzo Abe, who has a good prosodic personality in his speeches. To synthesize natural-sounding and prosodically rich speech, accurate phrasing, robust duration prediction, and rich intonation modeling are important. For these purpose, we propose pause position prediction based on conditional random fields (CRFs), phone-duration prediction using random forests, and mora-based emphasis context labeling. We examine the effectiveness of the above techniques through objective and subjective evaluations.

Keywords

Parametric Speech Synthesis, Hidden Markov Model (HMM), Prosodic Personality, Prosody Modeling, Conditional Random Field (CRF), Random Forest, Emphasis Context

1. Introduction

In the near future, people will have their own personal robots that support their daily life by communicating each other. To achieve such robots, speech recognition and synthesis interfaces are indispensable to make the communication of human-machine close to that of human-human. Currently, the use of speech recognition and synthesis technologies is rapidly spreading in smartphones (e.g., iPhone Siri), information guide in public facilities, and automotive navigation systems. Speech synthesis is a technology for generating speech from a text, and recently statistical parametric approach [1] based on hidden Markov models (HMMs) [2] has been widely studied and used [3]. However, most of the studies focus on synthesizing reading-style speech of news articles where the speaking style is always stable without prosodically rich expressions such as emphasis and emotions. Prosody of speech generally represents accent, intonation, rhythm, power, and phrasing (pause insertion) and has a rich personality. As a next step of speech synthesis to generate more human-like speech for various applications including humanoid robots, synthesizing speech with a rich prosodic personality is an essential issue.

In this paper, authors propose novel techniques for adding a rich personality to synthetic speech using a framework of HMM-based speech synthesis and machine learning. One of the final goals of this study is to achieve synthetic speech of Japanese prime minister, which gives sufficient impact and demands in practical applications. Speeches of the current prime minister, Shinzo Abe, are available in internet movies such as messages to the Japanese people and world leaders which are officially provided from the government. The speeches are very different from reading-style speech and contain prosodically rich expressions to emphasize important points and not to make audience bored. To achieve such more human-like speech synthesis with a limited amount of celebrity speech, the following techniques are presented in this paper.

- Prediction of phrase breaking based on conditional random fields (CRFs)
- Robust prediction of phone durations using random forests
- Speech parameter generation with emphasis context based on a mora unit to preserve rich intonation of natural speech

In most of the speech synthesis research, the phrasing information, *i.e.*, the positions of pause insertion, is manually given. However, the pause position sometimes strongly depends on the target speaker and, we need to automatically predict the positions from an input text in practical applications. In the speeches of Abe, he often inserts many pauses to clearly pronounce each word or phrase, and this style is very different from a general reading style. To model and predict the positions of phrase breaking, we use CRFs as label sequence modeling. For the duration modeling, hidden semi-Markov models (HSMMs) [4] are used for explicit modeling of state duration distribution [5]. However, the prediction accuracy of phone durations decreases when a sufficient amount of training data is not available. To improve the accuracy, the authors introduce phone-duration prediction using random forests [6] which is a kind of ensemble training [7]. Finally, speech parameter generation with mora-based emphasis context is presented to preserve rich intonation of natural speech, which is a variation of quantized fundamental frequency (F0) context [8] used also in voice conversion [9] and very low bit-rate speech coding [10].

The rest of this paper is organized as follows: In Section 2, we introduce a brief overview of parametric speech synthesis based on HMMs, which is a baseline speech synthesis system in this study. Section 3 describes speech materials used in this study. The role of prosody in speech synthesis and the problem of training data limitation are also explained in the section. Then, Section 4 explains the details of the proposed techniques to improve the prosodic personality when the training data of the target speaker is limited. In Section 5, the proposed techniques are compared with the baseline system through objective and subjective experiments and the results are discussed. Section 6 summarizes this study and refers to the future work.

2. Parametric Speech Synthesis Based on HMMs

In the HMM-based speech synthesis, speech parameter sequences, e.g., spectral and F0 features, are modeled in phone units as is the same as the case of HMM-based speech recognition. The advantage of the HMM-based speech synthesis compared to traditional concatenative speech synthesis is to generate smooth and stable speech parameters by considering dynamic features with a relatively smaller amount of speech data. Different from speech recognition, the modeling of prosodic features, *i.e.*, F0 and duration, is indispensable in speech synthesis. Since F0 has no value in silence and unvoiced regions, a special treatment is necessary such as the use of F0 interpolation [11] and multi-space probability distribution HMMs (MSD-HMMs) [12]. In the acoustic modeling, the acoustic property of speech parameters is affected by not only the current phoneme but also various factors such as preceding and succeeding phoneme, accent, stress, and sentence length. To take these factors into account, the factors are used as contexts and context-dependent HMMs are trained. Since the number of the combinations of contextual factors is too large, the contexts are tied using decision-tree-based context clustering [13] in the model training, and the number of unique contexts is reduced. In the phase of speech synthesis, a speech parameter sequence is generated based on a maximum likelihood criterion using the constraint of static and dynamic features [14]. Finally, a speech waveform is synthesized using a vocoding tool.

3. Speech Materials with a Rich Prosodic Personality

3.1. Speeches of the Japanese Prime Minister Abe

The HMM-based speech synthesis, which is a baseline in this study, is a corpus-based approach. This means that we need to prepare the speech data of a target speaker. The target speaker in this study is Shinzo Abe who is the 97th prime minister of Japan and is one of the most famous person in Japan. Since it is impossible to recording his voice in a standard way, the authors use speech data that is available at video hosting services. The type of the speeches is messages to the Japanese people at the annual events such as ones for Tohoku earthquake and official comments to the world leaders. However, the total length of collected speech data that have acceptable quality for speech synthesis is only about six minutes. We discarded utterances that included noise, reverberation, and unclear pronunciation in advance.

In a typical HMM-based speech synthesis, we prepare speech samples and corresponding texts, and make labels with phone boundary information. However, some utterances of Abe were very long, and automatic phone segmentation sometimes failed.

To avoid the problem, we divided a long utterance into short utterances based on pause. As a result, we had 319 utterances where 260 utterances included no pause. These utterance were used in the experiments of Section 5.

3.2. Pause Insertion for Voice Personality

Although pause insertion (phrase breaking) is used for breathing, it is also used to control speaking rate intentionally, to catch an attention, and to give calm impression to listeners. The position of pause insertion depends strongly on speakers, and the number of pause insertions also differs depending on speakers. Table 1 compares the average numbers of pauses per a minute between Japanese professional narrators with a reading style and the prime minister Abe. The two male and two female narrators are included in ATR Japanese speech database [15] set B. From the table, we found that Abe uses phrase breaking much more than the narrators. This result indicates that a general phrase breaking rule from an input text will degrade the voice personality of synthetic speech and intended effect appearing in the original speech is not always communicated to listeners correctly.

3.3. Role of Intonation and Speech Rate in Personality

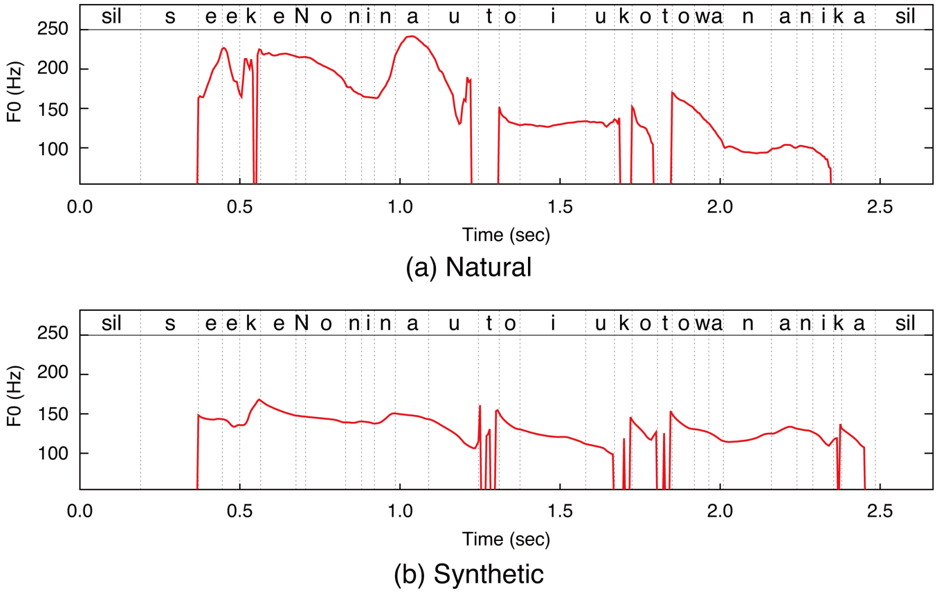
Precise prediction of speech intonation plays a crucial role in communicating para-linguistic information as well as improving naturalness of synthetic speech. Speech having clear intonation with emphasis expressions enables a speaker to make the listener understood the key point of the utterance. However, most of the speech synthesis systems cannot model and predict emphasis expressions automatically, and the synthetic speech loses such para-linguistic expressions. Figure 1 shows an example of natural and synthetic speech samples of Abe. It is found that the natural speech has a clearer F0 curve than synthetic speech. Specifically, there is a clear peak of the F0 pattern around 1.0 sec in natural speech. However, such feature disappears in synthetic speech, and the F0 pattern become flattened. This example indicates that the quality of the synthetic speech would be improved if we can model emphasis expressions in the model training.

3.4. Problem of Training Data Limitation

As is described in Section 3.1, the amount of speech data of Abe obtained from the internet is very limited. Although HMM-based speech synthesis can synthesize speech using a smaller amount of speech data of a target speaker than concatenative speech synthesis with unit selection, we typically need more than several tens of minutes

Table 1. Comparison of average pause insertion counts per minute between the prime minister Abe and professional narrators.

	Professional narrator				
Speaker	MHT	MMY	FTK	FKS	P.M. Abe
Count	23	24	18	17	30



(a) Natural

(b) Synthetic

Figure 1. Comparison of F0 contours between natural and synthetic speech samples.

training data to synthesize acceptable quality in naturalness. A straightforward way for this problem is to use speaker adaptation techniques such as maximum likelihood linear regression (MLLR) [16] with an average voice model [17] that is an acoustic model trained using speech data of multiple speakers. However, the adaptation performance depends on the average voice model, and the performance of the adaptation from a reading-style average voice model to a different type of speech, e.g., spontaneous speech, is not always satisfactory [18]. Therefore, the authors do not use the combination of the average voice model and a speaker adaptation technique in this study.

4. Prosodic Personality Improvement with Limited Data

In this section, the authors propose three techniques to improve the prosodic personality of synthetic speech when the amount of speech data of the target speaker, *i.e.*, the prime minister Abe in this study, is limited. Specifically, positions of pause insertion are predicted from an input text using CRFs. The accuracy of predicting phone durations is also improved by using random forests as an ensemble training technique. Furthermore, emphasis context based on a mora unit is introduced which can be automatically obtained by using differential features between natural and generated F0 parameter sequences. These techniques enable a TTS system to represent personal prosodic characteristics close to those of Abe while maintaining naturalness of synthetic speech.

4.1. Overview of the Proposed Speech Synthesis System

Figure 2 shows the outline of our text-to-speech system including three proposed techniques explained in the following sections. The system is named *Abe-droid* speech synthesis system[1]. In the figure, the boxes highlighted in yellow indicate the proposed

[1] The name comes from android which is a kind of humanoid robot.

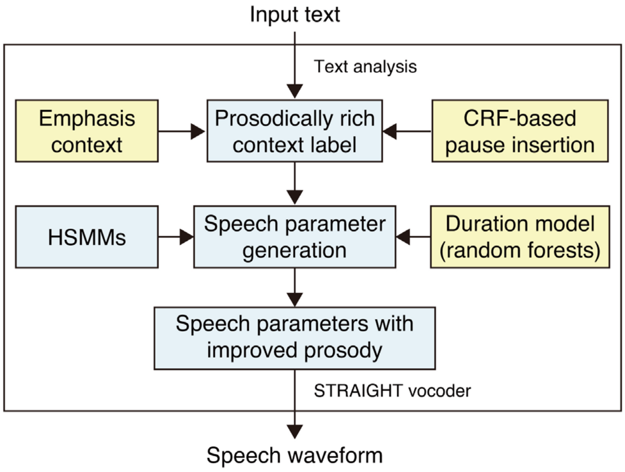

Input text
Text analysis
Emphasis context
Prosodically rich context label
CRF-based pause insertion
HSMMs
Speech parameter generation
Duration model (random forests)
Speech parameters with improved prosody
STRAIGHT vocoder
Speech waveform

Figure 2. Overview of the synthesis part of the proposed text-to-speech system. The proposed techniques are highlighted in yellow.

techniques in this paper, *i.e.*, CRF-based prediction of pause insertion position, robust duration modeling using random forests, and the use of mora-based emphasis context. When an input text is given, the text is converted to a context-dependent label sequence which has prosodically rich representation. At this time, pauses are automatically inserted based on a CRF-based pause insertion model, which is explained in Section 4.2. Emphasis context is also added to the labels. The emphasis context for the training data is automatically obtained using differential features. The detail is explained in Section 4.4. Then, speech parameter is generated using context-dependent HSMMs and prosodically rich context labels. The duration of each phone is determined using the duration model based on random forests (Section 4.3). Finally, a speech waveform is synthesized using a vocoder such as STRAIGHT [19].

4.2. Estimation of Pause Position Based on CRFs

In this study, pause positions are modeled and predicted using CRFs [20]. CRFs are used for a problem of sequence labeling where an appropriate label sequence y, e.g., part of speech tags, is predicted when an input sequence x is given. Let F be a sequence of features. $\varphi_f(x,y)$ stands for how many times a feature $f \in F$ appears in the set of (x,y), $\Phi(x,y)$ denotes its vector representation. The importance of each feature is represented by weight θ_f that is a parameter of a CRF, and Θ is its vector representation. Then, the conditional distribution for a CRF is given by

$$P(y|x)=\frac{\exp\langle\Theta,\Phi(x,y)\rangle}{\sum_{f\in F}\exp\langle\Theta,\Phi(x,y)\rangle} \tag{1}$$

where

$$\langle\Theta,\Phi(x,y)\rangle=\sum_{f\in F}\theta_f\varphi_f(x,y). \tag{2}$$

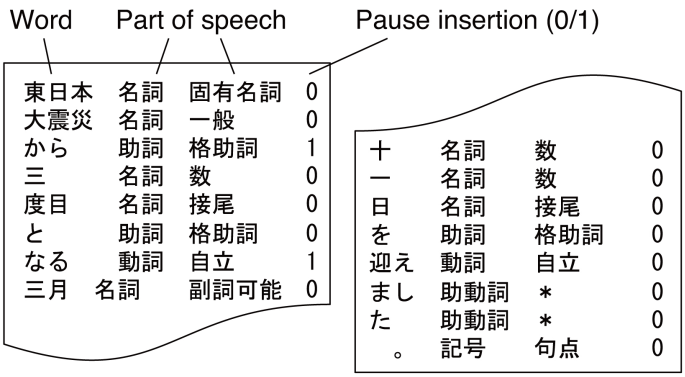
A set of model parameters is determined based on the maximum likelihood criterion.

To apply CRFs to Japanese text, we use a tool of Japanese morphological analysis, MeCab [21]. MeCab outputs surface form, part of speech (POS), subdivided POS 1, subdivided POS 2, subdivided POS 3, conjugated form, conjugation type, base form, reading, and pronunciation. In this study, we use only three factors, surface form, POS, and subdivided POS 1. For the surface form, preceding and succeeding forms are taken into account as well as the current form. Similarly, for the POS and subdivided POS 1, two preceding and two succeeding forms are taken into account in addition to the current ones. Figure 3 shows an example of the created training data in Japanese. The binary flags in the fourth field represent whether a pause is inserted after the morpheme or not.

4.3. Robust Phone-Duration Prediction Using Random Forests

Phone is the smallest unit of speech where we can distinguish the sound. Phone durations in an utterance are strongly related to local and global tempo and rhythm of speech. Therefore, modeling and predicting phone durations precisely are very important because they affect various properties of speech, e.g., speech naturalness, speaker individuality, speaking style, and emotional expression. In a preliminary experiment, we examined the performance of duration modeling in two ways. The first technique is to use standard HSMMs where state-duration distributions are explicitly modeled by Gaussian probability density functions (pdfs). This is a sophisticated way but has been shown to be worse than using an external duration model [22]. Therefore, we also used an external tree-based duration prediction model as the second technique where the distributions of phone durations are modeled as Gaussian pdfs and the model parameters are tied using a single context-dependent decision tree.

Both techniques work well when a sufficient amount of training data is available. However, the condition of this study is very severe and the training data is very limited, *i.e.*, only about six minute data is available. In that case, more robust prediction approach is required. We use random forests for this purpose. A random-forest technique is one of the machine learning techniques based on ensemble training and was


Word
Part of speech
Pause insertion (0/1)

```
東日本  名詞  固有名詞  0
大震災  名詞  一般      0
から    助詞  格助詞    1
三      名詞  数        0
度目    名詞  接尾      0
と      助詞  格助詞    0
なる    動詞  自立      1
三月  名詞    副詞可能  0
```

```
十    名詞    数      0
一    名詞    数      0
日    名詞    接尾    0
を    助詞    格助詞  0
迎え  動詞    自立    0
まし  助動詞  *       0
た    助動詞  *       0
  。  記号    句点    0
```

Figure 3. Example of training data (in Japanese) for CRF to predict pause position.

applied to speech synthesis for spectral parameter prediction [23]. **Figure 4** shows the outline of the proposed duration prediction using random forests. In the training phase, we make N subsets of training data by random sampling and construct a decision tree for each subset. These trees are used in the synthesis phase. An input text is converted to a context-dependent label sequence and is inputted into respective decision trees. Then, median filtering is applied to the output durations, and finally we obtain a predicted duration. When the number of subsets is even, two median values are obtained and we use the mean of the values as a predicted duration.

4.4. Intonation Control Using Mora-Based Emphasis Context

As is described in Section 3.4, modeling and synthesizing expressive speech that has a variety of local expressions is difficult when using a standard HMM-based speech synthesis framework. This is because the context labels used in model training and speech synthesis have no information of such local variations. For this problem, we proposed a prosody enhancement technique based on differential features of F0 and quantization [24] to capture the emphasis expressions in accent phrases of Japanese speech. To achieve more precise prediction of expressive speech such as speeches of Abe, similarly to the previous study, we here propose automatic mora-based emphasis expression labeling for training data. Mora is a basic unit for pronunciation in Japanese language and has similar characteristics to syllable in other languages, e.g., English. Japanese is a language of pitch accent, and we control an accent by changing relative pitch of each mora in an accent phrase.

The mora-based emphasis labeling is achieved as follows. First, standard HSMMs are obtained using training data without emphasis labels. We once generate F0 parameter sequences for the training sentences using the trained HSMMs. When comparing generated and natural F0 sequences, there are large differences in the region of speech having emphasis expressions. Hence, we calculate the differences between generated

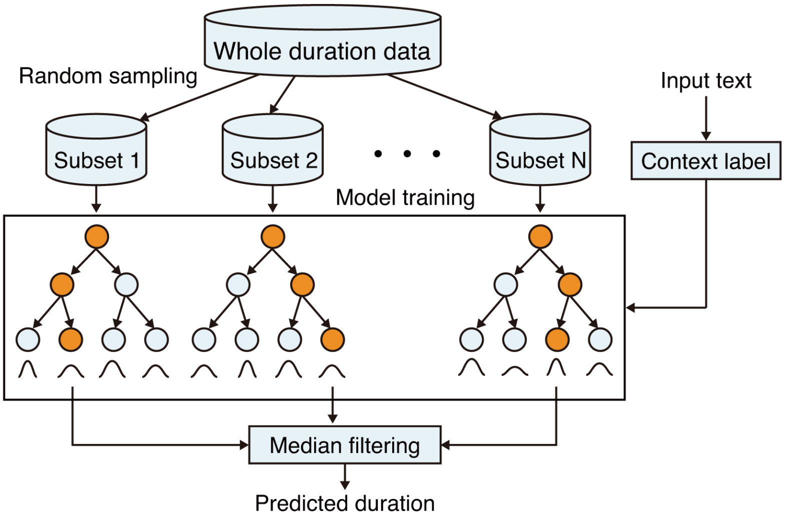


Figure 4. Phone duration modeling and prediction using random forests.

and natural F0 sequences and quantize the values into three levels, high (positive emphasis: 1), neutral (no emphasis: 0), and low (negative emphasis: −1), for each mora unit. The process is summarized as follows:

1) Train context-dependent HSMMs using conventional labels with only linguistic information.

2) Generate F0 sequences from the training sentences using the HMMs obtained above.

3) Calculate average log F0 values, f_o and f_s, of natural and synthetic speech for each mora unit.

4) Calculate the average log F0 difference $d = f_o - f_s$.

5) Classify d into three classes: a) $d < -\alpha$ (low), b) $-\alpha \le d < \alpha$ (neutral), and c) $d \ge \alpha$ (high), where positive value α is a classification threshold.

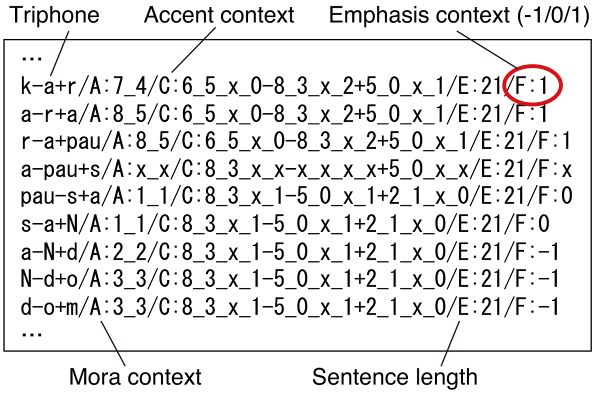
The threshold for the quantization can be automatically optimized using training data [24]. **Figure 5** shows an example of context-dependent labels including emphasis context that is automatically obtained for the training data. In the figure, triphone is shown in the left filed, accentual factors are shown in the center field, and emphasis context is shown in the right field. From the figure, we found that a mora sequence/ ara/has positive emphasis (1) and /Ndo/ has negative emphasis (−1).

5. Experiments

In this section, we incorporated our prosody modeling techniques described in Section 1 into the conventional baseline HMM-based speech synthesis and compared the performance through objective and subjective evaluations. In the objective evaluations, the prediction accuracy of pause positions, phone durations, and intonation similarity are examined. In the subjective evaluations, naturalness and similarity of synthetic speech are evaluated with five-point scale tests.

5.1. Experimental Conditions

We used about six-minute speech data of Abe that was described in Section 3.1. The



```
…
k-a+r/A:7_4/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1
a-r+a/A:8_5/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1
r-a+pau/A:8_5/C:6_5_x_0-8_3_x_2+5_0_x_1/E:21/F:1
a-pau+s/A:x_x/C:8_3_x_x-x_x_x_x+5_0_x_x/E:21/F:x
pau-s+a/A:1_1/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:0
s-a+N/A:1_1/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:0
a-N+d/A:2_2/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1
N-d+o/A:3_3/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1
d-o+m/A:3_3/C:8_3_x_1-5_0_x_1+2_1_x_0/E:21/F:-1
…
```

Figure 5. Example of a context-dependent label sequence with emphasis context. The symbol x means that there is no definition of the corresponding context.

total number of utterances was 319. 300 utterances were used as training data, and remaining 19 utterances were used as test data. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis [19] was used to extract spectral envelope, F0, and aperiodicity features with a five ms frame shift. The spectral envelope was converted to mel-cepstral coefficients using a recursion formula. The aperiodicity features were converted to average values for five frequency sub-bands, *i.e.*, 0 - 1, 1 - 2, 2 - 4, 4 - 6, and 6 - 8 kHz. The resultant feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, log F0, five average band aperiodicities, and their delta and delta-delta coefficients. The total number of dimensions was 138. We used five-state left-to-right HSMMs with no skip between states. Each state had a single Gaussian pdf with a diagonal covariance matrix. In the decision-tree-based context clustering, minimum description length (MDL) was used as a stopping criterion. In the baseline system, triphone, mora position, accent information, and sentence length were used as contextual factors.

5.2. Accuracy of Predicted Pause Insertion with CRFs

First, we evaluated the performance of predicting the positions of pause insertion based on CRFs. Table 2 shows a confusion matrix of predicted and correct classes of pause insertion. From the table, we found that more than 92% of the pause positions were correctly predicted using CRFs. In a practical application, listeners will perceive the prediction error as unnatural only when pauses are incorrectly inserted. This indicates that only 3.4% of pauses inserted by CRFs can affect the speech naturalness. In this experiment, the prediction accuracy is good even though the amount of training data is very limited. One of the reasons for this result is that the speeches were official messages and Abe regularly inserted pauses into the utterances. Therefore, we might need more data to achieve sufficient accuracy of pause insertion when the target speaker is a person who is inexperienced at speaking officially, which is our future work.

5.3. Effect of Random Forests in Phone-Duration Prediction

Next, we evaluated the effectiveness of using random forests in phone-duration prediction. The number of subsets for the random forests was set to six, and each fifty utterances were used to construct decision trees using context clustering with an MDL-based stopping criterion. For comparison, duration prediction techniques using HSMMs and a single tree were also evaluated. Root mean square (RMS) error of phone durations between natural and synthetic speech was used as an objective measure. Table 3

Table 2. Ratio (%) of classified boundaries for pause insertion.

		Correct class	
		Pause	Not pause
2*Predicted class	Pause	14.9	3.4
	Not pause	4.6	77.1

Table 3. Comparison of RMS errors (ms) of phone durations.

HSMM	Single tree	Random forests
64.73	66.39	**36.87**

shows the result. From the table, we found that the use of random forests in phone-duration prediction substantially decreased the objective distortion and made the phone durations closer to those of the natural speech when compared to the conventional techniques.

To investigate the detail of the effect, we also examined the distributions of predicted phone durations with the conventional and proposed techniques. Figure 6 shows the histograms of phone durations. From the figure, we found that the phone-duration prediction based on random forests reduced the RMS errors of durations more than 100 ms compared to the conventional techniques. In addition, the distribution of phone durations in random forests is closer to a Gaussian distribution than the other techniques, which indicates that phone durations were well modeled and predicted by using random forests.

5.4. Effect of Emphasis Context for Intonation Improvement

We also examined whether the use of emphasis context improves the intonation of synthetic speech. For the quantization of differential F0 features, we first determined threshold α using training data. The objective measure of F0 similarity to the natural speech is the RMS error of log F0 between natural and synthetic speech. For the threshold optimization, threshold was changed from 0.0 to 1.0 with an increment of 0.1, and the smallest value, $\alpha = 0.12$, was used as the optimal threshold. Then, emphasis contexts of high, neutral, and low, were determined for the training and test utterances. For comparison, we trained HSMMs in three conditions. The first was the conventional HSMMs without emphasis context. The second and the third were HSMMs with emphasis context using the initial threshold of $\alpha = 0.0$ and the optimal threshold $\alpha = 0.12$, respectively. The RMS errors of log F0 (cent) were calculated between natural and synthetic speech. Table 4 shows the result. From the table, it is seen that the use of emphasis context substantially reduced the F0 distortions. We also found that the distortions were further reduced by the threshold optimization Figure 7 shows an example of F0 contours with and without emphasis context. From the figure, it is seen that the F0 contour generated with emphasis context is closer to natural speech and has a clearer intonation than that without emphasis context.

5.5. Total Subjective Evaluation

Finally, we conducted total subjective evaluation tests to examine the effect of each proposed technique for improving prosody in synthetic speech generated from limited training data. We evaluated speech synthesis in five different conditions as follows:

Baseline Conventional HSMM-based speech synthesis

Pau-predict Baseline with CRF-based pause insertion

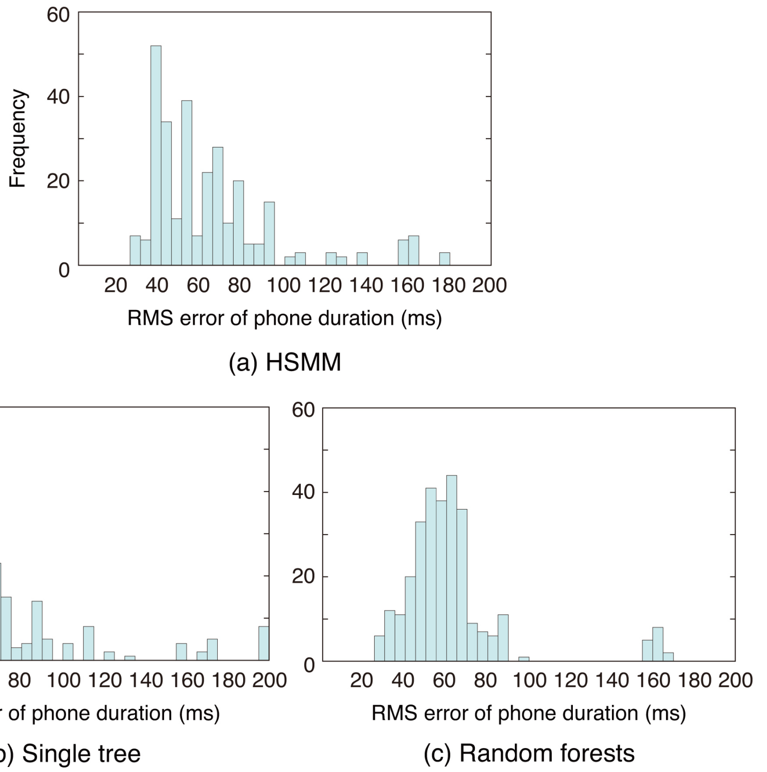
Pau-correct Baseline with correct pause insertion

Pau + dur Pau-correct with phone-duration prediction using random forests

Pau + dur + emph Pau + dur with emphasis context

For all synthetic speech samples, the pause length was set to 0.65 (sec) which was the mean value of the pauses included in the training data. The participants were ten native Japanese speakers. We evaluated the naturalness and similarity of the synthetic speech samples with mean opinion score (MOS) tests. For the similarity test, participants listened to natural speech samples as reference before the synthetic speech stimuli. Naturalness and similarity were evaluated on a five-point scale: "1" for bad, "2" for poor, "3" for fair, "4" for good, and "5" for excellent. During the MOS tests, participants could repeat to play sentences to evaluate the utterances as many times as required. Figure 8 shows the average scores for the respective techniques.

From the figure, we found that naturalness and similarity of the baseline system is



(a) HSMM

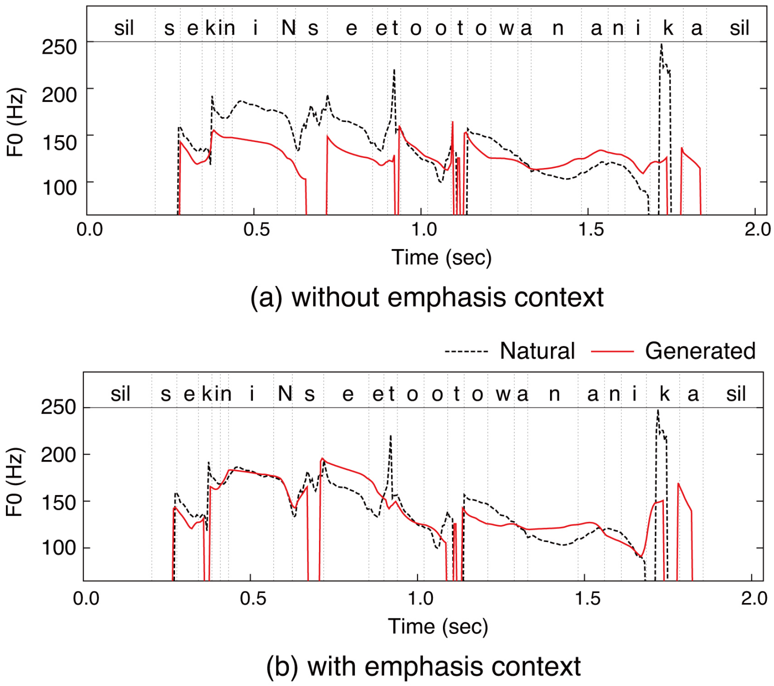
(b) Single tree

(c) Random forests

Figure 6. Histograms of predicted phone durations in the conventional and proposed techniques.

Table 4. Effect of emphasis context with an optimized threshold when comparing RMS errors (cent) of F0 for test data.

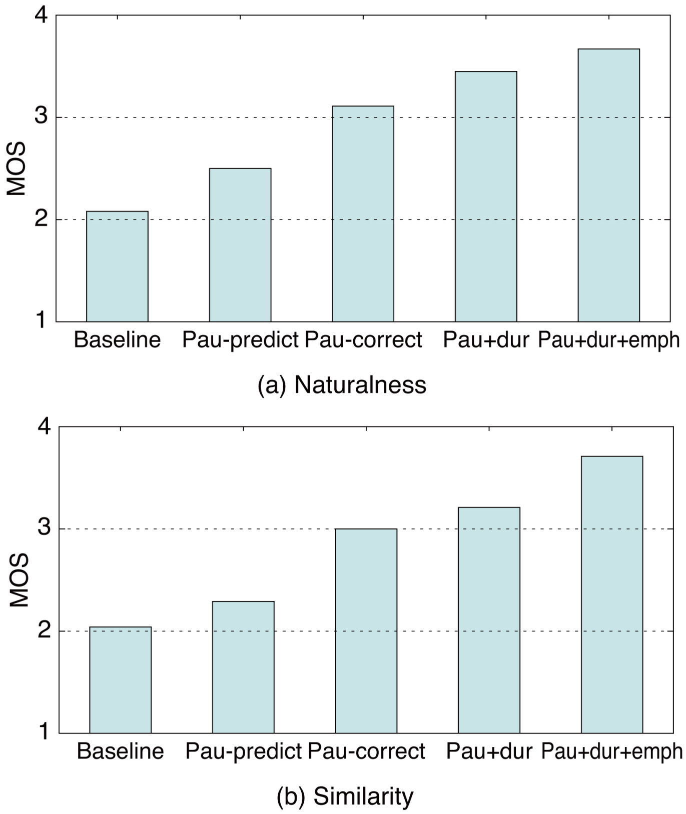
2*HSMM	Emphasis context	
	Default	Optimal
	($d = 0.0$)	($d = 0.12$)
413.2	285.3	**249.3**



(a) without emphasis context

(b) with emphasis context

Figure 7. Effect of the proposed emphasis context in terms of F0 contours.



(a) Naturalness

(b) Similarity

Figure 8. Results of subjective evaluation of synthetic speech in different conditions.

not satisfactory when the amount of the target speaker is very limited and the speech is prosodically rich. By introducing pause prediction, there was 0.5 point improvement in the naturalness evaluation, and similarity was also improved. However, we found that the prediction performance was still insufficient when comparing Pau-predict and Pau-correct. One of the reasons of this gap is that some of the test utterances were relatively long and included many pauses, and the naturalness and similarity degraded even when one pause was incorrectly inserted. The proposed duration prediction and emphasis modeling worked well and both of them improved naturalness and similarity.

6. Conclusion

The final goal of this study is to achieve a speech synthesis interface that is commercially valuable and has a rich personality. For this purpose, we focused on synthesizing the voice of the prime minister of Japan, Shinzo Abe, as the target speaker. We proposed techniques for HMM-based speech synthesis to achieve an interface of prosodically rich speech synthesis when the target speaker is a celebrity but the available speech data is limited. We presented CRF-based prediction of the position of pause insertion, robust phone-duration prediction using random forests, and the use of emphasis context for mora units. The objective and subjective evaluation results have shown that all techniques improved the performance of speech synthesis from the baseline HMM-based speech synthesis system. The current sysmtem has a limitation that the emphasis context must be added manually to the input text for synthesis, and hence the automatic labeling of emphasis context for test data is our future work. In addition, we will attempt to introduce speaker adaptation technique under the condition that multiple speakers' speech data are available in advance. Synthesizing emotional speech of celebrities is also an remaining task.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP15H02720 and Step-QI school in department of electrical, information and physics engineering, Tohoku University.

References

[1] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1999) Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. *European Conference on Speech Communication and Technology*, 2347-2350.

[2] Rabiner, L.R. and Juang, B.-H. (1986) An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, **3**, 4-16. https://doi.org/10.1109/MASSP.1986.1165342

[3] Zen, H., Tokuda, K. and Black, A. (2009) Statistical Parametric Speech Synthesis. *Speech Communication*, **51**, 1039-1064. https://doi.org/10.1016/j.specom.2009.04.004

[4] Levinson, S. (1986) Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. *Computer Speech & Language*, **1**, 29-45. https://doi.org/10.1016/S0885-2308(86)80009-2

[5] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (2007) A Hidden

Semi-Markov Model-Based Speech Synthesis System. *IEICE Transactions on Information and Systems*, E90-D, **825-834**. https://doi.org/10.1093/ietisy/e90-d.5.825

[6] Liaw, A. and Wiener, M. (2002) Classification and Regression by Randomforest. *R News*, **2**, 18-22.

[7] Dietterich, T.G. (2000) Ensemble Methods in Machine Learning. *Proc. International Workshop on Multiple Classifier Systems*, 1-15.

[8] Nose, T., Ota, Y. and Kobayashi, T. (2010) HMM-Based Voice Conversion Using Quantized F0 Context. *IEICE Transactions on Information and Systems*, **E93-D**, 2483-2490. https://doi.org/10.1587/transinf.E93.D.2483

[9] Nose, T. and Kobayashi, T. (2011) Speaker-Independent HMM-Based Voice Conversion Using Adaptive Quantization of the Fundamental Frequency. *Speech Communication*, **53**, 973-985. https://doi.org/10.1016/j.specom.2011.05.001

[10] Nose, T. and Kobayashi, T. (2012) Very Low Bit-Rate F0 Coding for Phonetic Vocoders Using MSD-HMM with Quantized F0 Symbols. *Speech Communication*, **54**, 384-392. https://doi.org/10.1016/j.specom.2011.10.002

[11] Yu, K., Thomson, B. and Young, S.J. (2010) From Discontinuous to Continuous F0 Modelling in HMM-Based Speech Synthesis. *Proceedings of* 7*th ISCA Speech Synthesis Workshop*, Kyoto, 22-24 September 2010, 94-99.

[12] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T. (2002) Multi-Space Probability Distribution HMM. *IEICE Transactions on Information and Systems*, **E85-D**, 455-464.

[13] Riley, M. (1990) Tree-Based Modelling for Speech Synthesis. *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, 25-28 September 1990, 229-232.

[14] Tokuda, K., Kobayashi, T. and Imai, S. (1995) Speech Parameter Generation from HMM Using Dynamic Features. 1995 *International Conference on Acoustics, Speech, and Signal Processing*, Detroit, 9-12 May 1995, 660-663. https://doi.org/10.1109/ICASSP.1995.479684

[15] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K. (1990) ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. *Speech Communication*, **9**, 357-363. https://doi.org/10.1016/0167-6393(90)90011-W

[16] Leggetter, C.J. and Woodland, P.C. (1995) Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, **9**, 171-185. https://doi.org/10.1006/csla.1995.0010

[17] Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T. (2001) Text-to-Speech Synthesis with Arbitrary Speaker's Voice from Average Voice. 7*th European Conference on Speech Communication and Technology*, Scandinavia, 3-7 September 2001, 345-348.

[18] Koriyama, T., Nose, T. and Kobayashi, T. (2010) Conversational Spontaneous Speech Synthesis Using Average Voice Model. 11*th Annual Conference of the International Speech Communication Association*, Chiba, 26-30 September 2010, 853-856.

[19] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A. (1999) 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*, **27**, 187-207. https://doi.org/10.1016/S0167-6393(98)00085-5

[20] Lafferty, J., McCallum, A. and Pereira, F.C. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 18*th International Conference on Machine Learning*, Williamstown, 28 June-1 July 2001, 282-289.

[21] Kudo, T. (2005) Mecab: Yet another Part-of-Speech and Morphological Analyzer. https://github.com/taku910/mecab

[22] Latorre, J., Buchholz, S. and Akamine, M. (2010) Usages of an External Duration Model for HMM-Based Speech Synthesis. 5*th International Conference on Speech Prosody*, Chicago, 11-14 May 2010, 1-4. http://speechprosody2010.illinois.edu/papers/100073.pdf

[23] Black, A.W. and Muthukumar, P.K. (2015) Random Forests for Statistical Speech Synthesis. *Proceedings of Interspeech*, Dresden, 6-10 September 2015, 1211-1215.

[24] Maeno, Y., Nose, T., Kobayashi, T., Koriyama, T., Ijima, Y., Nakajima, H., Mizuno, H. and Yoshioka, O. (2014) Prosodic Variation Enhancement Using Unsupervised Context Labeling for HMM-Based Expressive Speech Synthesis. *Speech Communication*, **57**, 144-154. https://doi.org/10.1016/j.specom.2013.09.014