論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Statistical Speech Synthesis Using Extended Context and Gaussian Process Regression
著者(和文)	郡山知樹
Author(English)	Tomoki Koriyama
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9323号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:小林 隆夫,羽鳥 好律,伊東 利哉,小池 康晴,篠崎 隆宏
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9323号, Conferred date:2013/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Statistical Speech Synthesis Using Extended Context and Gaussian Process Regression

Tomoki Koriyama

September 2013

Summary

This thesis describes novel approaches for synthesizing speech with prosodic variability and naturalness. There are a variety of applications of speech synthesis and there have been increasing demands for such applications. Although the variability and naturalness of synthetic speech have been improved, the ability of generating natural sounding speech is still insufficient. This thesis focuses on spontaneous conversational speech that has much prosodic variability. The purpose of this study is to improve the variability using spontaneous speech data by realizing the framework that can synthesizes natural-sounding spontaneous conversational speech.

First, extended context is introduced to synthesize natural-sounding spontaneous conversational speech with prosodic variability in the hidden-Markov-model-based speech synthesis framework. Several context sets that can be obtained from the Corpus of Spontaneous Japanese are introduced and the effectiveness of the context sets are evaluated. The results of objective evaluation show that the phone prolongation and tone labels are effective for improving generated F0 and duration. It has been confirmed that the synthetic speech using extended context offers more natural-sounding speech than conventional contexts from the subjective evaluation.

Next, prosodic-event-based HMM (prosodic-unit HMM) is proposed to improve the naturalness of prosody of spontaneous conversational speech. The modeling unit proposed prosodic-event-based HMM is the segment between two tone labels that represents prosodic events such as pitch falling by accent or pitch rising of boundary pitch movement (BPM). The proposed HMM is expected to reduce the model parameters of F0 because there are less prosodic events derived from F0 features than phones that strongly depends on spectral features. The results show that the proposed technique gives a

more compact model and more variation in generated F0 than phone-unit HMM.

The prosodic variability and naturalness of synthetic speech is improved by extended context and prosodic-event-based HMM. However the naturalness of spectral features is still insufficient. Then, a speech synthesis framework based on Gaussian process regression is proposed to improve the naturalness of spectral features. Block-based sparse GP approximations such as local GPs and PIC are used for trajectory modeling of utterances with feasible computation. Moreover, for the generation of smooth parameter trajectory, frame context including nearby phone information and its kernel is defined. From the objective and subjective evaluation, the proposed method using the PIC approximation and the extended context achieved better performance than the HMM-based methods.

Acknowledgments

First, I would like to express my thanks to Professor Takao Kobayashi, Tokyo Institute of Technology, for all of his support, encouragement, and guidance. Also, I would like to thank to Professor Yoshinori Hatori, Professor Toshiya Ito, Professor Yasuharu Koike, and Professor Takahiro Shinozaki of Tokyo Institute of Technology for their kind suggestions.

I would like to also thank Dr. Takashi Nose of the Kobayashi Laboratory at Tokyo Institute of Technology. His substantial help in my work are deeply appreciated. Over the years, I have benefited greatly from interaction with members of the Kobayashi, Shinozaki, Ida, and Sumita Laboratories. There are too many people to mention individually, but I must thank Prof. Junichi Yamagishi (currently with National Institute of Infomatics) and Dr. Takashi Masuko (currently with Toshiba Corporation).

Finally, I would like to give my special thanks to my family for all their support over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General background

Voice is one of the most important tools of communicating with human beings. The sound wave of voice can convey not only linguistic information but also emotions and dialog acts, called para-linguistic information, and characteristics of speakers. Speech synthesis is a technique of generating the voice artificially using computers, and this enables computers to speak to humans. Nowadays, speech synthesis has been used in a variety of fields.

One of the most important features of speech synthesis is to convey arbitrary text information to people without visual texts or images. Historically, text-to-speech (TTS) has been used for a screen reader for people with visual impairments. Speech synthesis embedded in automobile navigation systems offers road information with voice to drivers who are required to look around.

Another application of speech synthesis is a tool that supports human-human communication. In a medical field, speech synthesis techniques are employed as alternative voice output communication aids (VOCAs) for patients whose speech become disordered. For example, it revealed that a speech synthesis technique enables personalized VOCAs that have patients' original voice characteristics in [1].

A recent interesting field is an application for audio and video contents creation. Singing voice synthesis, closely related to speech synthesis, has been widely used in musical creation [2] along with the popularization of

video sites such as YouTube. In addition, the state-of-the-art application of speech synthesis is the synthesis of audiobooks [3], [4]. The advantages of using synthesis may include the low cost without sound recording and the flexibility of editing synthetic voices.

With a spread of domains, the requirements for speech synthesis system have increased. One of the requirements is the realization of prosodic variability. Prosody of spontaneous speech that appears in human-human conversation could enrich VOCA system. Prosody included in actors' performance could enhance the quality of audio contents. Also, naturalness is required for intelligibility and pleasantness of synthetic speech. In general, if we collect huge training data, we can synthesize natural-sounding speech. However, it is not easy to prepare huge data of specific domains. Therefore, it is preferable to synthesize speech with a small amount of training data.

Statistical parametric speech synthesis [5], including hidden-Markov-model-based (HMM-based) speech synthesis [6], has been developed as a technique that is expected to achieve the naturalness and variability of synthetic speech using a relatively small amount of data. In statistical parametric speech synthesis, speech signals are converted to low-dimensional speech parameters like mel-cepstrum and fundamental frequency (F0), which we refer to as acoustic features. We statistically model the relationship between the acoustic features and phonetic and prosodic variable factors, called *contexts*, such as phonemes, syllables, accents and other factors that mainly derives from transcriptions or input texts. And, we generate acoustic features from input contexts.

The statistical parametric model can provide a compact and flexible representation of speech. This characteristic has been utilized to change speakers characteristics and speaking styles by changing their parameters. For example, speaker adaptation with average voice model based on HMM enabled training using only a small amount of target speaker's data by changing the mean and variance parameters of output probability distribution functions (pdfs) of HMMs [7]. For the expressive speech synthesis, parametric representation makes it easy to model multiple emotions and speaking styles by the style-mixed model with the global context of style types [8]. Although the variability and naturalness have been improved, the ability of generating

human-like clear speech is still insufficient.

1.2 Scope of thesis

This thesis describes a novel approaches to improving the variability and naturalness of synthetic speech in the statistical speech synthesis.

In chapter 3, first, we focus on the improvement of variability using spontaneous conversational speech data. Since the spontaneous conversational speech has diverse prosodic variability including boundary pitch movements (BPMs) and filled pauses. However, it is difficult to model these factors in conventional contexts used for reading-style speech. We extend the contexts based on the annotated data of the Corpus of Spontaneous Japanese (CSJ) [9]. The categories of extended context include phone prolongation, utterance style, tone label, disfluency, complementary phoneme, word, and clause. By adding the contextual factors of these categories to conventional context set, we evaluate which category is important to express the prosodic variability. Since adding too many contexts often causes a over-fitting problem, we attempt to choose an effective context subset. Also, in order to alleviate over-fitting problem on training using many contextual factors, we intdroduce a new criterion for context clustering, the minimum number of observations.

Next, in chapter 4, we attempt to improve the prosody model of spontaneous speech. We propose a new modeling unit of HMM-based speech synthesis based on prosodic events which correspond to folding points of F0 movements. The proposed prosodic-event-based unit is the segment between the tone labels of X-JToBI [10]. We incorporate the proposed unit to the HMMs for F0, and use mora-normalzed position for the timing prediction of the labels

The prosodic variability and naturalness of synthetic speech have been improved by extended context and prosodic-event-based HMM. However the naturalness of spectral features is still insufficient. In chapter 5, we propose a novel framework of statistical speech synthesis based on Gaussian process regression [11]. The model of GPR is designed for directly predicting frame-level acoustic features from corresponding information on frame context that

is obtained from linguistic information. GPR-based speech synthesis can overcome some problems of HMM-based speech synthesis, e.g., the limited number of model parameters and the mismatch between discrete HMM-state space and continuously changing acoustic features.

Chapter 2

Statistical Parametric Speech Synthesis

This chapter describes parametric speech synthesis system using statistical models. Firstly, the overview of the system is provided. Then, HMM-based speech synthesis is introduced, which is one of the most widely-used techniques.

2.1 Overview of statistical parametric speech synthesis

Statistical parametric speech synthesis has grown in popularity in over the last years [6], [12]–[14]. Figure 2.1 outlines the system of statistical parametric speech synthesis. In the training stage, phonetic and prosodic contexts are extracted from the texts with annotation information. Also, frame-level acoustic features are extracted from speech samples using speech analysis. Acoustic features include spectral feature, fundamental frequency (F0), and duration information. Mel-cepstrum is often used as the spectral feature. After that, a statistical model that represents the relationship between acoustic features and corresponding contexts is trained. In the synthesis stage, the contexts of input text are extracted and an acoustic feature sequence is generated using the contexts and the trained model. Finally, output speech waveform is synthesized from the generated acoustic feature sequence. The

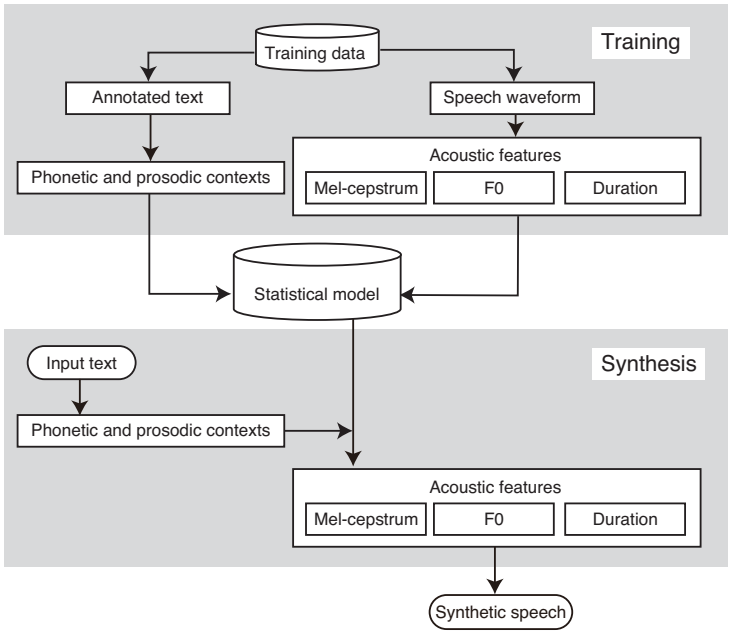


Figure 2.1: Outline of statistical parametric speech synthesis.

term "parametric" means that speech waveform is *parametrized* into acoustic features and the term "statistical" implies that the features are *statistically* modeled.

2.2 HMM-based speech synthesis

In HMM-based speech synthesis [6], HMM and decision trees is used as a statistical model. The contexts are defined for each phone and have phoneme, accent, part of speech, breath group, and utterance length information. As acoustic features, we use mel-cepstrum, log F0, and band aperiodicity [13], which include not only static features but also dynamic ones. In the training stage, context-dependent HMMs are trained. Since the combination of the context sets is huge, decision-tree-based context clustering is performed to predict the HMM parameters of unseen contexts. In the synthesis stage, HMMs are chosen using the decision tree and a parameter sequence are generated by maximum likelihood parameter generation (MLPG). The following subsections describe the details of HMM-based speech synthesis.

2.2.1 Model topologies for HMM-based speech synthesis

Hidden Markov model (HMM) is known to be an effective model to express time series observation. An HMM λ consists of state transition probability parameters $\{a_{ij}\}$ and observation probability functions $b_i(\mathbf{o}_t)$. Here, a_{ij} denotes the transition probability from the state i to the state j. $b_i(\mathbf{o}_t)$ represents the probability of observation of time t, $\mathbf{o}_t$, if the state if i. The likelihood of observation sequence $\mathbf{O} = (\mathbf{o}_1, \ldots, \mathbf{o}_T)$ is given by

$$\begin{aligned} P(\mathbf{O}|\lambda) &= \sum_{\mathbf{q}} P(\mathbf{q}|\lambda) P(\mathbf{O}|\mathbf{q}, \lambda) \\ &= \sum_{\mathbf{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t) \\ &= \sum_{\mathbf{q}} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\mathbf{o}_t) \end{aligned} \tag{2.1}$$

where $\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_T)$ is a hidden state sequence. The likelihood is computed effectively using forward-backward algorithm and the parameters are trained using Baum-Weltch algorithm, which is EM algorithm for HMM.

A left-to-right and no-skip HMM is commonly used for the model of speech parameter sequences because it can appropriately model the variability of acoustic features. A hidden semi-Markov model (HSMM) [15] has been proposed as a model that includes explicit state duration representation by Gaussian distributions, whereas the HMM implicitly expresses state duration by the transition probabilities.

Gaussian distribution is commonly used for the pdf of the observation. For modeling F0s, which is unobserved in the unvoiced region, by continuous pdfs, a multi-space probability distribution HMM (MSD-HMM) has been proposed [16]. In the MSD-HMM, the pdf is changed dependently on whether the frame is voiced or unvoiced, that is, a Gaussian distribution is used for a voiced region and a fixed value is used for an unvoiced region.

2.2.2 Context clustering

Context clustering means collecting similar phonetic and prosodic contexts [17]. There are two aims in context clustering. The first aim is to define the context-dependent HMM whose context is not included in training data (called *unseen context*) by using the model parameters of similar contexts. The second is to improve the reliability of model parameters. If the model parameters of a HMM are estimated by a single segment, the reliability of the parameters tends to be low. The context clustering enables reliable model parameter estimation using multiple segments included in the same cluster.

A decision tree is used for context clustering. Each leaf node is split using the question, e.g. whether the current phone is /i/ or not, that most increases the likelihood. As the stopping criterion of the node splitting, the minimum description length (MDL) determined by the number of model parameters and the amount of training data has been shown to be effective [18].

2.2.3 Maximum likelihood parameter generation from HMM

Maximum likelihood parameter generation (MLPG) [19] is a simple but an effective way of generating parameter sequence from HMMs. First, the optimal state sequence $\mathbf{q}^*$ is generated using the state duration means. After that, the most likely parameter sequence $\mathbf{C}^*$ is calculated as follows

$$\begin{aligned} \mathbf{C}^* &= \underset{\mathbf{C}}{\operatorname{argmax}}\, P(\mathbf{O}|\mathbf{q}^*, \lambda) \\ &= \underset{\mathbf{C}}{\operatorname{argmax}}\, P(\mathbf{WC}|\mathbf{q}^*, \lambda) \end{aligned} \tag{2.2}$$

where $\mathbf{W}$ is a window matrix that converts parameter sequence $\mathbf{C}$ to the acoustic feature vector sequence including dynamic feature, $\mathbf{O}$.

2.3 Statistical parametric speech synthesis in this thesis

In this thesis, we explore novel methods based on the framework of Fig. 2.1. In chapter 3, the phonetic and prosodic contexts are extended for sponta-

neous speech synthesis in the framework of the HMM-based speech synthesis. In chapter 4, we examine the HMM in the statistical model for modeling of F0 of spontaneous speech. Specifically, the unit of HMM is changed into a prosodic-event-based unit. In chapter 5, we change the statistical model from HMM and decision tree to Gaussian process regression. For the initial examination of the method based Gaussian process regression, we use phonetic information for the contexts and mel-cepstrum for the acoustic feature.

2.4 Conclusion

In this chapter, the basic overview of statistical speech synthesis system and HMM-based speech synthesis system is introduced. The basic components of HMM-based speech synthesis, HMM, context clustering, and MLPG, are described.

Chapter 3

Spontaneous Speech Synthesis using Extended Context

This chapter proposes an extended context set for generating the prosodic variability of spontaneous conversational speech in HMM-based speech synthesis. Since the conventional context set used for HMM-based reading-style speech synthesis is insufficient for conversational speech synthesis, we introduce extended contexts derived from the Corpus of Spontaneous Japanese. Using the stopping criteria for decision-tree clustering to alleviate over-fitting by increasing contexts, we compare conventional contexts and extended contexts, and show that the contexts about phone prolongation and X-JToBI tone tier label are effective. Furthermore, we examine the automatic prediction of a part of contexts for practical applications, and confirm the comparable naturalness by predicted contexts to that using true contexts.

3.1 Introduction

Although the quality of the synthetic speech of neutral reading-style has been improved and become closer to that of the natural speech, the quality is generally unsatisfactory when the conventional techniques are applied to the spontaneous and/or conversational speech synthesis. When a very large corpus of conversational speech is available, concatenative speech synthesis based on unit selection has been shown to be able to produce natural sound-

ing speech like a human [20]. Recently, there have been alternative attempts for spontaneous and/or conversational speech synthesis [21]–[24] using HMM-based synthesis which has shown its advantage in a relatively small amount of training data. In [21], fundamental frequency (F0) contours and phone durations were modeled based on the quantification theory type I. Another prosody modeling technique was proposed in [22], where state-based voice transformation from read speech was used. In [23], a technique based on the multi-space distribution HMM (MSD-HMM) [25], which is widely used for the F0 modeling in the HMM-based speech synthesis, was also evaluated. To reduce the required amount of spontaneous speech, an average-voice-based technique was shown to be effective [24].

Although the naturalness of the synthetic speech could be improved by using the above techniques, there is still a large acoustic difference between real and synthetic speech. One of the critical problems is degradation of prosodic variability. This is inevitable when we use the conventional context that was designed for the HMM-based speech synthesis of reading-style speech.

In this chapter, we incorporate additional context sets into the HMM-based speech synthesis framework to improve the prosodic variability of the generated spontaneous conversational speech. Newly introduced contexts are derived from the annotation data included in the Corpus of Spontaneous Japanese (CSJ) [9], which is a large-scale database designed for the study of the spontaneous speech. Several kinds of context sets are evaluated to examine the effectiveness of each context category. Furthermore, to avoid an over-fitting problem that occurs in the model training, we propose two types of stopping criteria for tree-based context clustering.

3.2 Extended context for spontaneous conversational speech

In this study, we examine 12 context categories shown in Table 3.1. These contexts can be determined automatically by the labels annotated in the CSJ core. In the conventional HMM-based speech synthesis, the information

Table 3.1: Context categories.

BASELINE		ADDITIONAL	
A	Phoneme	F	Phone prolongation
B	Mora	G	Utterance style
C	Accent	H	Tone label
D	Breath group	I	Disfluency
E	Utterance length	J	Complementary phoneme
		K	Word
		L	Clause

about phoneme, mora, accent phrase, breath group, and utterance length has been used as the *full context*. We refer to this context set as BASELINE. The details of respective additional context categories are as follows.

Phone prolongation: Phone prolongation is the phenomenon that vowel or consonant is uttered for a longer period than the ordinary and it often occurs in the utterances with thinking, surprising, or emphasizing. This is distinct from the lexical long vowels appearing in Japanese dictionaries. In CSJ, this phenomenon is labeled like "sugo<H>i (very)" or "kai<Q>seki (analysis)," where <H> and <Q> represent vowel and consonant prolongation, namely, "o" and "s" are prolonged, respectively.

Utterance style: In CSJ, some utterance styles are also added on the transcription texts when a certain mora is uttered with a particular style. We adopt three styles as the contexts related to mora information, namely, laughing, whisper, and uncertainly pronounced sound.

Tone label: Japanese is a pitch-accent language and one accent phrase is composed of several words. An accent phrase has one accent type that determines the relative pitch movement over the accent phrase. The relative pitch movement of Japanese reading-style speech can be well represented by the accent type. However, the pitch movements of spontaneous conversational speech are much more complicated than those

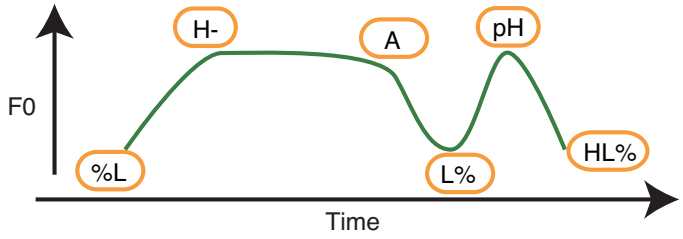


Figure 3.1: Schematic example of the relationship between X-JToBI tone tier labels and the F0 contour of an accent phrase which ends with rise-fall. Each inflection point is labeled.

Table 3.2: List of tone tier labels.

Label	Usage
%L	Phrase-initial boundary
H−	Phrase tone, the peak of phrase-initial rise
A	Accent, beginning of accentual fall
L%	End of accentual fall or phrase-final boundary of fall pattern
H%	Phrase-final boundary of rise pattern
LH%	Phrase-final boundary of fall-rise pattern
HL%	Phrase-final boundary of rise-fall pattern
HLH%	Phrase-final boundary of rise-fall-rise pattern
pL	Low tone pointer accompanying LH% and HLH%
pH	High tone pointer accompanying HL% and HLH%
FL	Filler-high
FH	Filler-low

of reading-style speech, and it is difficult to represent such a movement by the accent type only. One of the essential information for conversational speech is boundary pitch movements (BPMs) such as rise, fall, rise-fall, and fall-rise which occur in question, confirmation, and other speech acts. To take the complicated pitch movements including the BPM into account, CSJ uses the intonation labeling scheme of X-JToBI [10], the extension of ToBI. For instance, a tone tier defined

by X-JToBI has labels on the folding points of F0 contour as shown in Fig. 3.1. In this study, we use the type of pitch movement of accent phrase and the difference of positions between tone label and current mora.

Disfluency: Conversational speech has many disfluent utterances that interrupt the flow of speech and express affective content. There are three types of disfluency annotated in CSJ: filler, word fragment, and restatement. These disfluency of the phrases are used as the contexts.

Complementary phoneme: Precise phonetic information is also labeled in CSJ, that is, some kind of utterance parts which is hard to be categorized into general phoneme sets. In this study, tags <sv> and <cl> defined in CSJ are adopted as the contexts. <sv> means the vocal cord vibration after vowel, and <cl> denotes the burst which appears with explosion.

Word: Spontaneous conversational speech has peculiar phenomena about the morpheme information such as fusion, omission, and euphony of words because of the informality of spontaneous speech. The word information including such phenomena as well as part of speech is embedded to two types of word. "Short-unit word" approximately corresponds to a vocabulary entry of ordinary Japanese dictionary, and "long-unit word" is composed of a few of short-unit words.

Clause: The clause is a grammatical unit that consists of a subject and a predicate, and the clause boundaries are automatically determined by transcription data in CSJ. We use the clause type and the mora position in the clause as the contexts.

3.3 Stopping criteria for context clustering

In general, the minimum description length (MDL) has been shown to be effective [18] for the context clustering in HMM-based speech synthesis. However, when the extended context described in Sect. 3.2 is incorporated into the spontaneous conversational speech synthesis, the MDL criterion does not

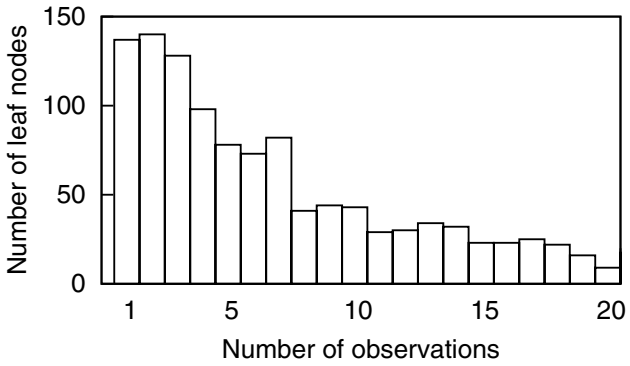


Figure 3.2: Histogram of the number of observations contained in one leaf node.

always work well. This is because the prosodic variation of such speech is much larger than that of the reading-style speech, as a result, over-fitting often occurs when the amount of training data is limited. Figure 3.2 shows a histogram of the number of observations contained in one leaf node of a decision tree. The decision tree was constructed using about 22.5 minutes data of a female speaker (ID=19) included in the CSJ database. From the figure, we can see that there are many leaf nodes that have only a few observations.

To alleviate the over-fitting problem, we attempt to use the minimum occupation count or the minimum number of observations. The minimum occupation count is a parameter that restricts the total number of observation frames in each leaf node as the stopping criteria. However, this criterion might not work well since spontaneous conversational speech includes a lot of phone prolongations and the number of frames of an observation segment sometimes becomes much larger. In such a case, the criterion based on the minimum number of observations which restricts the total number of observation speech samples in each leaf node would be more suitable.

3.4 Objective evaluation

3.4.1 Experimental conditions

We conducted evaluation experiments using conversational speech data of two female speakers (ID=19 and 514) included in the CSJ database. Each

speaker was non-professional speaker and uttered three sets of conversational speech—two interviews and a task-oriented dialog. The total length of speech samples of each speaker was approximately 25 minutes. Speech signals were sampled at a rate of 16kHz. The STRAIGHT analysis [26] was used for extracting the spectral envelope and F0. The feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient and log F0, and their delta and delta-delta coefficients. We used hidden semi-Markov model (HSMM) [15] which has explicit duration distributions. The model topology was 5-state left-to-right context-dependent HSMM without skip paths. Each state had a single Gaussian distribution with a diagonal covariance matrix. For training and testing, the phonetic and prosodic context labels were automatically converted from the labels given in CSJ. Ten-fold cross-validation tests was performed in the evaluations.

3.4.2 Extended context

To evaluate the effectiveness of the extended context, the average distortions of generated spectrum, F0, and mora duration of synthetic speech were calculated against those of the original speech. Figure 3.5 shows the average mel-cepstral distance, root mean square (RMS) errors of log F0 and mora duration, respectively. In this case the minimum occupation counts were fixed to 5.0 but the minimum number of observations was not limited. In the figure, BASELINE represents the conventional context. ALL is the context set where all of the context categories described in Sect 3.2 are included in addition to BASELINE. It is seen that RMS errors of log F0 and mora duration were decreased significantly by using the extended context along with the conventional context set. On the contrary, there was no significant difference of the mel-cepstral distance between BASELINE and ALL.

To examine the effect of respective context categories, different context sets were evaluated where one context category was chosen from categories F to L of Table 3.1 and was added to BASELINE. The results in Fig. 3.5 indicate that the use of tone label (H) decreased the most. This means that the tone information such as inflection point of F0 curve and boundary pitch movement was important for the generation of a natural sounding log F0

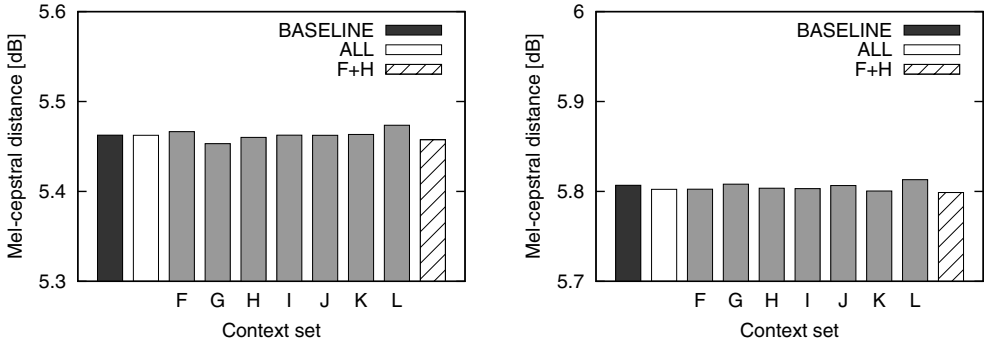


Figure 3.3: Spectral distortions with different context sets.

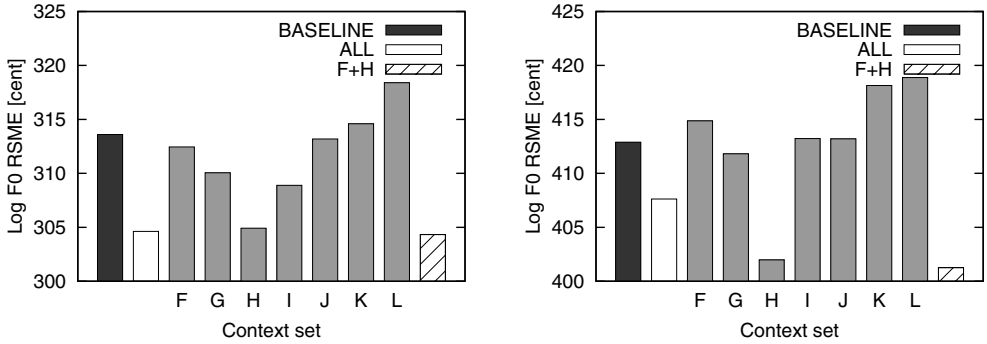


Figure 3.4: F0 distortions with different context sets.

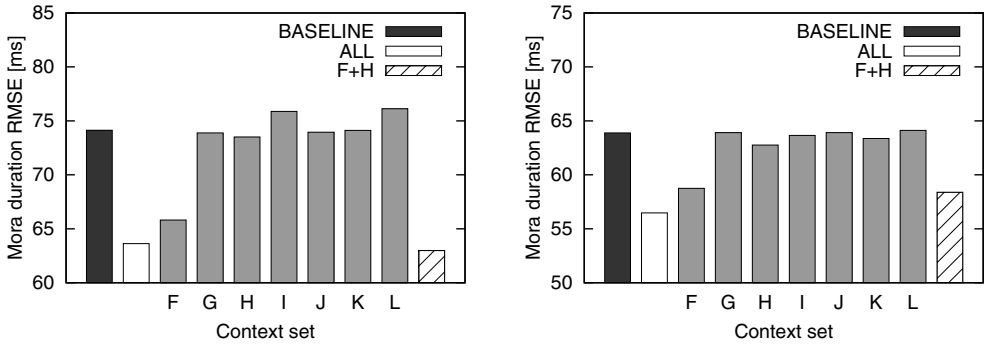


Figure 3.5: Duration distortions with different context sets.

pattern. Utterance style (G) and Disfluency (I) decreased the RMS errors slightly. In contrast, the F0 distortion increased when the context categories of word(K) or clause(L) were used. A possible reason is that an over-fitting occurred in the model training because of the insufficient training data. As for the mora duration, phone prolongation (F) worked well whereas the others

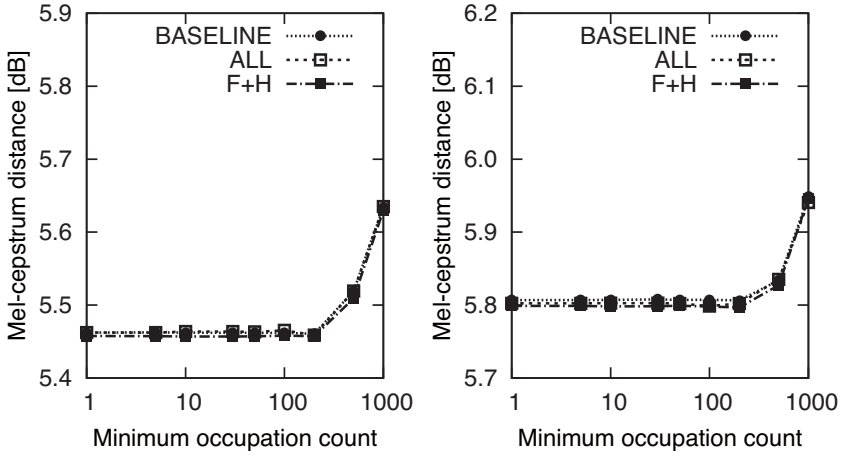
were not effective for the distortion reduction. In consideration of above results, another context set was also evaluated where the categories F and H were added to BASELINE. The results are shown as F+H in Fig. 3.5. We can see that the distortions of mel-cepstrum and duration of F+H were comparable to ALL. Moreover, the F0 distortion of F+H was slightly lower than that of ALL.

3.4.3 Stopping criteria

The effectiveness of the use of stopping criteria based on the minimum occupation count and the minimum number of observations was objectively assessed. The distortions of the acoustic features of synthetic speech were calculated against those of the original speech. We changed the thresholds of criteria for mel-cepstrum and log F0 features. Figures 3.6, 3.7, and 3.8 show average mel-cepstral distances and RMS errors of log F0 and phone durations. The minimum number of observations is not limited in Fig. 3.6, and the minimum occupation count is fixed to 5.0 in Fig. 3.7. The relationship between the leaf node size and the stopping criteria are shown in Figs. 3.9 to 3.11.

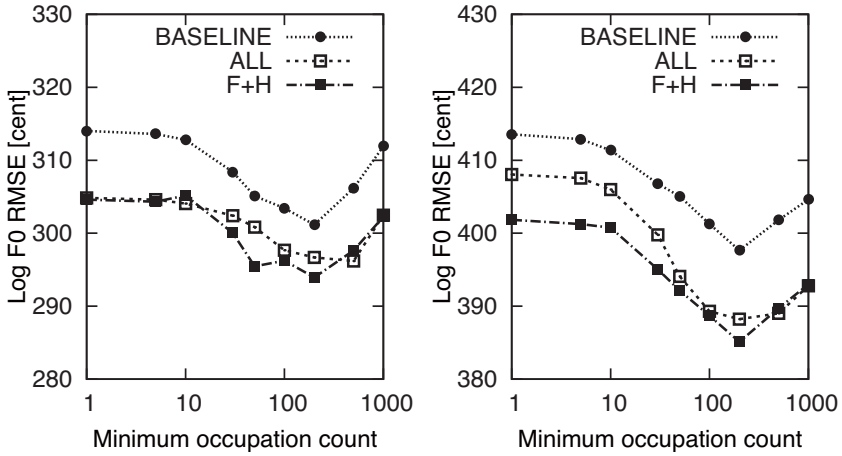
As for the mel-cepstral distance, it was not sensitive to the stopping criteria when the minimum number of observations was less than 100 or the minimum occupation count was less than 200. From these results, the stopping criteria appears to be not necessary for the mel-cepstrum. On the other hand, the F0 distortion decreased when the minimum occupation count or minimum number of observations were taken into account. This implies that the over-fitting problem was alleviated by introducing these stopping criteria into the clustering.

To examine the effect of the combinational use of two stopping criteria, the log F0 distortions for the speaker (ID=19) with the context set F+H were calculated when both of the criteria were used in the clustering. The result is shown in Fig. 3.12. From the figure, we see that the use of either one of the two criteria seems to be enough to suppress the over-fitting. When the results of the F0 distortion with two criteria are compared in Figs. 3.6 and 3.7, the criterion based on the minimum number of observations was less



(a) Spectral distortion (speaker #19)

(b) Spectral distortion (speaker #514)

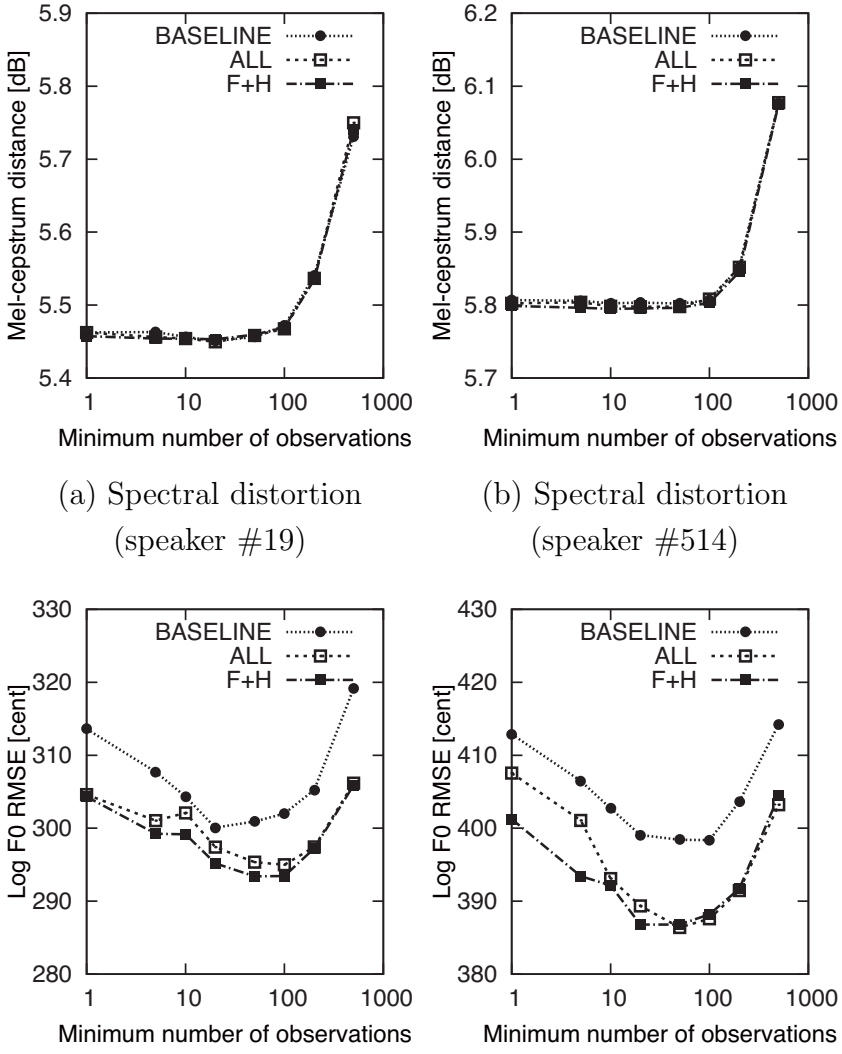


(c) F0 distortion (speaker #19)

(d) F0 distortion (speaker #514)

Figure 3.6: Spectral and F0 distortions as a function of the minimum occupation count.

sensitive to distortion variation than that based on the minimum occupation count.



(a) Spectral distortion (speaker #19)

(b) Spectral distortion (speaker #514)

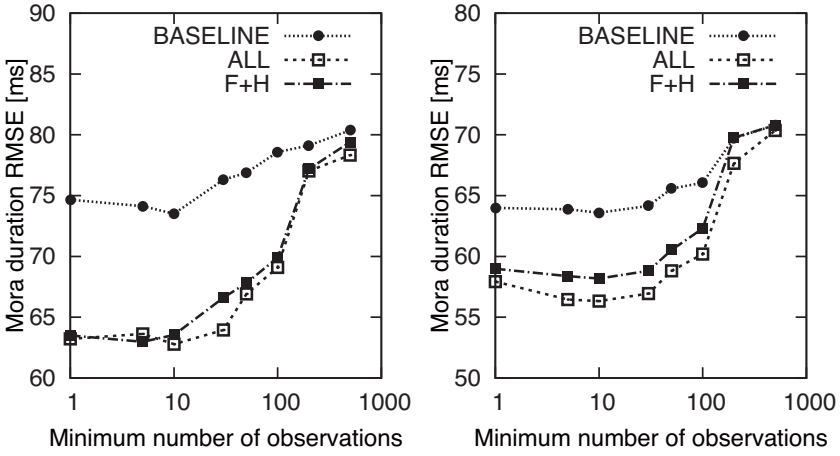
(c) F0 distortion (speaker #19)

(d) F0 distortion (speaker #514)

Figure 3.7: Spectral and F0 distortions as a function of the minimum number of observations.

3.5 Position prediction of tone labels

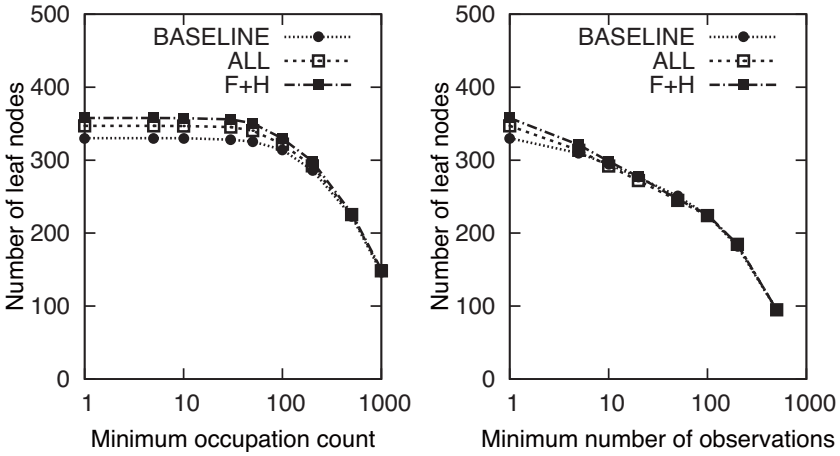
As described in Sect. 3.4.2, incorporation of the extended contexts reduced the distortions of acoustic features. However, it is unrealistic that users input all the extended contexts in the synthesis step. Even though it is ideal to determine all the extended contextual factors automatically, this is substan-



(a) Duration distortion (speaker #19)

(b) Duration distortion (speaker #514)

Figure 3.8: Mora duration distortions as a function of the minimum number of observations.

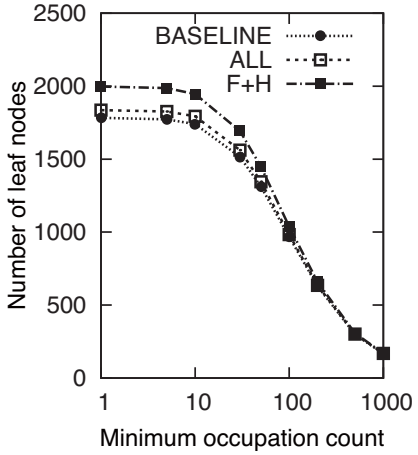


(a) Tree for mel-cepstrum using minimum occupation count

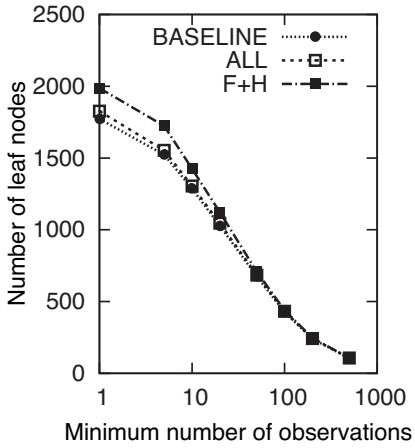
(b) Tree for mel-cepstrum using minimum number of observations

Figure 3.9: Relationship between tree size for mel-cepstrum and stopping criteria.

tially impossible because some contextual factors are not dependent on input text but other factors such as situations and dialog acts. Here we consider a practical framework where users modify some extended contexts manually

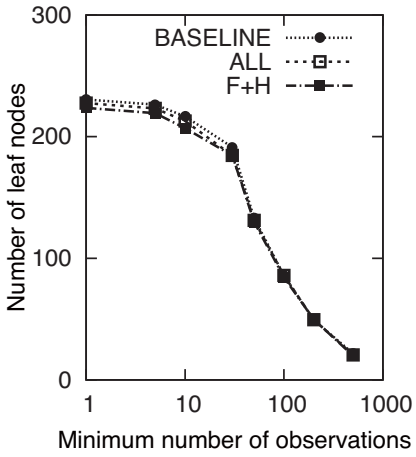


(c) Tree for log F0 using minimum occupation count



(d) Tree for log F0 using minimum number of observations

Figure 3.10: Relationship between tree size for log F0 and stopping criteria.



(e) Tree for state duration using minimum number of observations

Figure 3.11: Relationship between tree size for duration and stopping criteria.

based on the context set F+H, whose RMS errors of F0 and duration were comparable with that using all contexts. The context set includes the phone prolongation, the relative mora position of X-JToBI tone tier labels, and the type of pitch movement of accent phrase. It might not be complicated that

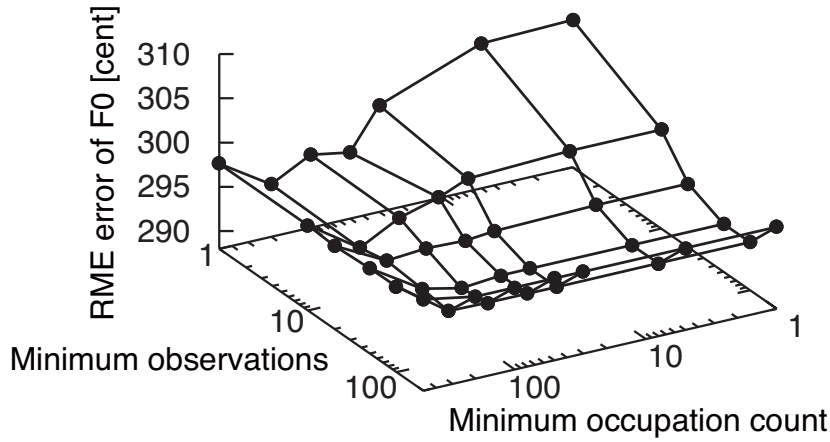


Figure 3.12: F0 distortions as a function of the minimum occupation count and the minimum number of observations.

they modify the contexts of the phone prolongation and the type of pitch movement of accent phrase, because they have only to add symbols that represent prolonged phones, boundary tones, prosodic fillers, and prosodic word fragments. On the other hand, in order to use the relative position as contextual factors, it is necessary to prepare the existence and position information of tone labels, which is difficult to give manually. Hence, we examine the method of determining the label position information automatically. In this section, we examined two types of determining methods: rule-based method and prediction using a decision tree.

3.5.1 Determination of label position by rule

The positions of X-JToBI tone tier labels can be determined roughly without using real F0 contours. For example, a phrase tone "H−" is generally located around the second or third mora of the accent phrase. The labels "L%" and "H%" in a rise tone phrase are usually positioned around the beginning and end of the last mora, respectively. Therefore, the rule listed in Table 3.3 is introduced to determine the label positions. The existence of the labels is determined by the following rule based on the fundamental meanings of the labels. If the accent type is 1, "H−" is nonexistent. If the accent type is 0, "A" is existent. The accuracy of the positions by the rule is also shown in

Table 3.3: Rule and accuracy of label positions [%]

Label	Position	Accuracy[%]
Phrase-initial boundary (%L)	Initial mora	88.51
End of phrase-initial rise (H−)	Second mora	63.96
Beginning of accentual fall (A)	Accent mora	86.80
End of accentual fall (L%)	Final mora	83.96
Phrase-final boundary (*%)	Final mora	89.71
Low tone pointer (pL)	Final mora	99.98
High tone pointer (pH)	Final mora	99.76
Filler (FL or FH)	Central mora	97.84

Table 3.3.

3.5.2 Determination of label position by decision tree

In this chapter, we examine a decision-tree-based method, C4.5, for prediction. C4.5 was implemented by WEKA [27]. The input variables consisted of the type of pitch movement and the conventional context set of accent phrase, including the length of phrase and the accent type. The output variable were defined by the relative position like +1, +2, −1, −2 of each label to each three kind of basic position—phrase initial, perceived accent mora, and phrase final. Ten-fold cross-validation was performed in the same way as the previous section.

Table 3.4 shows the accuracy of position information for each label and each basic position. The highest accuracy (indicated in bold) varies depending on the labels. From the table, it can be seen that the accuracy of "%L" and "H−" located around phrase initial was highest when the basic position was phrase initial, and that the accuracy of "*%" and high/low pointers located around phrase final was highest when the basic position was phrase final. Compared with the case using the rule, the accuracy of "H−" was improved largely. This could be explained by insufficient rule for "H−".

Table 3.4: Accuracy of predicted label positions [%]

Label / Reference position	Itinial	Accent	Final
Phrase-initial boundary (%L)	**90.45**	90.19	89.80
End of phrase-initial rise (H−)	**77.58**	77.58	76.72
Beginning of accentual fall (A)	88.59	**88.64**	87.77
End of accentual fall (L%)	82.84	80.89	**83.97**
Phrase-final boundary (*%)	89.28	86.71	**89.74**
Low tone pointer (pL)	99.52	99.45	**99.93**
High tone pointer (pH)	99.62	98.86	**99.73**
Filler (FL or FH)	**97.80**	97.80	97.68

3.5.3 Synthesis using prediction

Based on above results, we synthesized utterances using the extended context obtained by the rule and decision-tree-based prediction for the label position. The total length of speech samples of each speaker was approximately 25 minutes and 10-fold cross-validation was performed in the same way as Sect. 3.4. Based on the results of Sect. 3.4.3, the minimum number of observations for each leaf node in acoustic modeling was set to 50.

Tables 3.5 and 3.6 show acoustic distortions of F0 and mora duration between original and synthetic speech. The context set "F+H CORRECT" represents the case using annotated labels for synthesis, namely, it is identical to "F+H" in Sect. 3.4. "F+H W/O POSITION" does not use the label position information. "F+H RULE" and "F+H CORRECT" correspond to the methods where the label position information is determined by the rule and the prediction model, respectively.

When comparing "F+H CORRECT" with "F+H W/O POSITION," we can find that the label information is effective for decreasing distortions of F0 and mora duration. Moreover, "F+H RULE" and "F+H PREDICTED," which automatically determines the label position information, gave similar mora duration distortion to "F+H CORRECT" in speaker #19, and decreased RMSE of F0 by approximately 5 cents from "F+H W/O PO-

Table 3.5: F0 distortions using predicted positions of tone labels

	RMS error of F0 [cent]	
Context set	Speaker #19	Speaker #514
F+H CORRECT	293.4	386.8
F+H RULE	298.4	390.9
F+H PREDICTED	299.3	391.3
F+H W/O POSITION	300.9	396.0
BASELINE	300.9	398.4

Table 3.6: Mora duration distortions using predicted positions of tone labels.

	RMS error of mora duration [ms]	
Context set	Speaker #19	Speaker #514
F+H CORRECT	63.1	58.6
F+H RULE	62.9	58.6
F+H PREDICTED	62.9	58.8
F+H W/O POSITION	65.9	58.7
BASELINE	74.0	64.0

SITION." There were slight differences between "F+H RULE" and "F+H PREDICTED." From the results, we can find it important to

3.6 Evaluation of naturalness

The naturalness of the synthetic speech was evaluated for the four context sets: BASELINE, ALL, F+H CORRECT, and F+H PREDICTED by a MOS test. The minimum number of observations was set to 50 in ALL, F+H CORRECT, and F+H PREDICTED based on the above results. Ten Japanese participants listened to synthetic speech samples and rated the speech naturalness in a five point scale, i.e., 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. Each participant evaluated 20 utterances for each context set, randomly chosen from synthetic speech samples which were used for objective evaluation and were the utterances consisted of 10 or more moras.

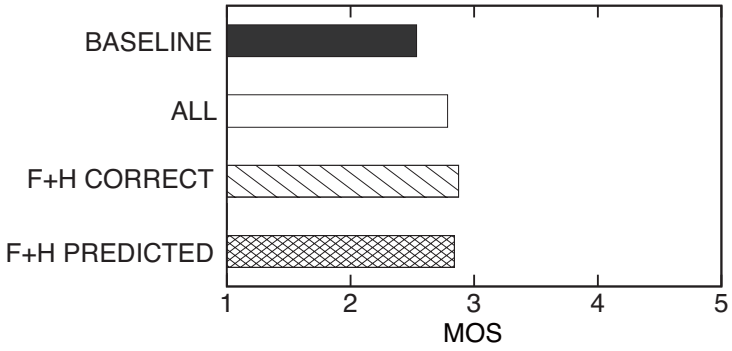


Figure 3.13: Results of a MOS test on naturalness of synthetic speech.

The average scores are shown in Fig. 3.13. ALL, F+H CORRECT, and F+H PREDICTED gave higher performance than BASELINE, and the difference is statistically significant at the 5% level. F+H CORRECT was comparable to ALL, and this indicates that the information of tone and phone prolongation is critical for the extended context in terms of the naturalness of the synthetic speech. Furthermore, F+H PREDICTED, which determines label position information automatically, gave comparable score to F+H CORRECT.

3.7 Conclusion

To synthesize spontaneous conversational speech with more prosodic variability, we have investigated the effectiveness of several context categories based on annotations of CSJ. We have also examined the stopping criteria of decision-tree context clustering to alleviate the over-fitting problem which comes from increasing contexts and conversational variability. The objective and subjective experiments showed that the reproducibility and naturalness of synthetic speech is improved by adding the contexts of phone prolongation and X-JToBI tone tier labels and by introducing the minimum number of observations for each leaf node of the decision tree. For the future work, it is important to generate the extended contexts automatically from the concept, speech act of speech, and other utterance information for practical conversational speech synthesis system.

Chapter 4

Prosodic-event-based HMM

In this chapter, we propose prosodic-event-based HMM for effectively modeling F0 pattern of spontaneous conversational speech in HMM-based speech synthesis. The prosodic-event-based HMM uses the segment such as pitch falling by accent or pitch rising of boundary pitch movement (BPM) as a modeling unit of HMM. The proposed HMM is expected to reduce the model parameters of F0 because there are less prosodic events derived from F0 features than phones that strongly depends on spectral features. We performed the objective and subjective experiments using spontaneous conversational speech data, and the results show that the prosodic-event-based HMM can significantly reduce the number of model parameters while keeping the quality of the synthetic speech.

4.1 Introduction

In the previous chapter, the naturalness of the synthetic speech of spontaneous conversation with much prosodic variability was improved by incorporating a set of extended prosodic contexts, such as phone prolongation and tone information, to HMM-based speech synthesis. However, the prosody generation of multiple speaking styles included in spontaneous speech is still insufficient. Here we focus on the modeling unit of HMM-based speech synthesis. In the HMM-based speech synthesis, fundamental frequency (F0) is usually modeled using phone-unit-based HMMs and trained synchronously

with spectral features. To model both voiced and unvoiced regions of the F0 pattern consistently, multi-space distribution HMM (MSD-HMM) [25] is utilized. Although this HMM is good at modeling of prosodic features of phone unit, it is not always suited for the F0 pattern of spontaneous speech well. This is because the positions of prosodic events such as accent and boundary pitch movement (BPM) do not always match those of phonetic movements. For instance, in a segment outside of prosodic events called *connection* in rise/fall/connection model [28], F0 features do not change so much as prosodic events even if the segment contains several phones. On the other hand, in a segment at the rise-fall pitch movement, F0 moves largely even if the segment has only one phone.

To alleviate this problem, there have been proposed different approaches to the F0 modeling, e.g., the use of hierarchical structures [29] and the use of longer units [30]. In this study, we propose an alternative approach to modeling F0 contour efficiently using prosodic-event-based HMM units. More specifically, we use components of prosodic events, such as the segment of pitch falling by accent and pitch rising by BPM, as the modeling unit. Since the prosodic events of one phrase are less frequent than the changes of phonemes, the proposed unit is expected to reduce the number of model parameters of F0, which leads to robust parameter estimation. We examine the effectiveness of the proposed F0 modeling technique through both objective and subjective evaluation experiments.

4.2 F0 modeling based on prosodic events

4.2.1 Prosodic labels

In the proposed technique, we define a speech synthesis unit using prosodic label information related to F0 contours. In this chapter, we employ the X-JToBI labeling scheme [10] for the proodic labels. As described in the previous chapter, X-JToBI is an extension of J-ToBI [31] that is the Japanese version of ToBI [32]. Table 4.1 shows the labels used in X-JToBI. These labels include the timing information of the folding points of F0 contours. The type of label depends on the function in prosodic events. Phrasal tone and

Table 4.1: X-JToBI tone tier labels

label	function	abbr.
%L	beginning of phrase	s
H−	end of pitch rise	m
A	beginning of pitch fall by accent	a
%LA	joint label of %L & A	b
L%(LTBPM)	low tone before BPM	t
L%(FBT)	end of phrase with fall tone	l
H%	end of phrase with rise tone	h
HL%	end of phrase with rise-fall tone	i
LH%	end of phrase with fall-rise tone	j
HLH%	end of phrase with rise-fall-rise tone	k
pH	high pointer in BPM	p
pL	low pointer in BPM	q
FL	filler with low pitch	FL
FH	filler with high pitch	FH

accent consist of "H−" and "A," respectively. Boundary pitch movements consist of "*%," "pH," and "pL." "L% (FBT: final boundary tone)" and "L% (LTBPM: low tone of BPM)" are distinguished by the function; the former expresses the end of the accent phrase with falling tone, and the latter is the start of BPM. Ordinary accent phrases can be expressed by the label sequence which starts with %L followed by other labels according with the label network shown in Fig. 4.1 and ends with the final boundary tone, "*%." Other phrases are prosodic fillers and prosodic word fragments. A prosodic filler is a filled pause which does not have neither a pitch rise nor a local pitch fall anywhere. If a speaker stops in or starts from the middle of accent phrase, the phrase is treated as a prosodic word fragment. This phenomenon is caused by disfluency of spontaneous speech.

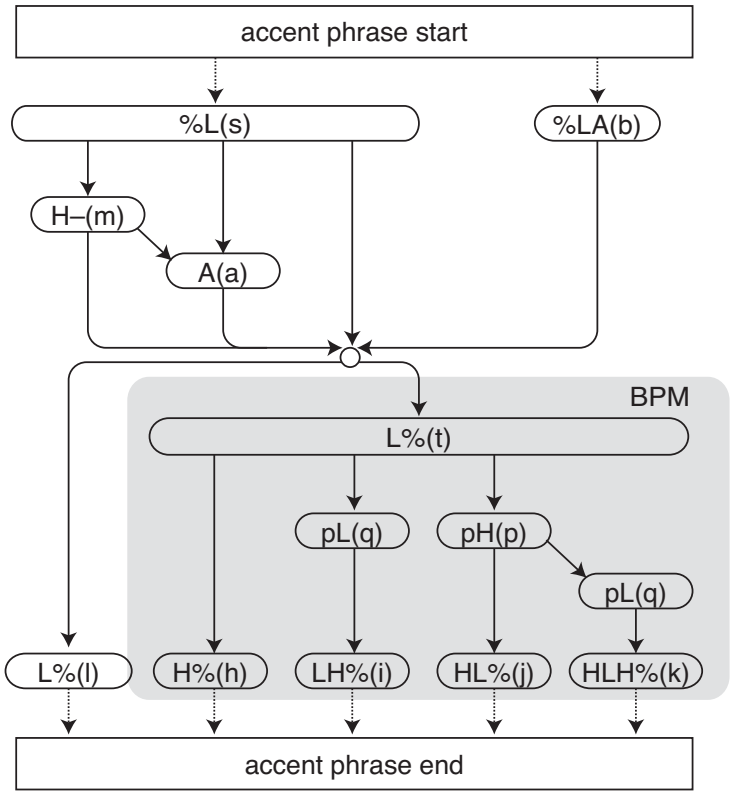


Figure 4.1: Network of prosodic labels

4.2.2 Prosodic-event-based HMM

The segment between X-JToBI tone tier labels can be regarded as the basic unit of a prosodic event. Hence, we adopt this segment as a unit of HMM for F0 modeling. We refer to the proposed prosodic-unit-based HMM as *prosodic-unit HMM*, whereas we refer to the conventional phone-unit-based HMM as *phone-unit HMM*. One prosodic unit is distinguished from others by the combination of the labels. For example, the segment between "%L" and "A" is labeled as "%L-A." To simplify the notation of prosodic unit, we use the combination of the single characters corresponding to the labels, which are shown in Table 4.1. The prosodic units are listed in Table 4.2. Here, "y" and "z" represent the beginning and end of the prosodic word fragments, respectively. The segment of filler is labeled as the label name, "FH" and "FL." "SP" means the prosodic space between accent phrases (e.g. the segment from "L%" to "%L,") and "PZ" and "SL" denote the pause and

Table 4.2: Prosodic units

type	unit names
normal segment	sm, sa, sl, st, ma, ml, mt al, at, bl, bt
BPM	th, tp, tq, pi, pq, qj, qk
prosodic filler	FH, FL
prosodic word fragment	yh, yl, az, bz, mz, yz
pause & silence	PZ, SL, SP

silence, respectively.

The training procedure of the prosodic-unit HMM is similar to that of the conventional phone-unit HMM. The value of log F0, its delta, and delta-delta coefficients are used as the features of HMM. To model voiced/unvoiced regions, we use MSD in a similar manner to phone-unit HMM. An HMM of each prosodic unit is initialized by the segmental K-means algorithm, and the parameters are refined by Baum-Weltch re-estimation. Then HMMs are clustered by their prosodic contexts. The context set for prosodic unit consists of quin-prosodic-unit and the information of the units of accent phrase, breath group, and utterance. By using the prosodic-unit HMM as the speech synthesis unit, we can model F0 patterns more efficiently with the prosodic label information compared to the case when using the conventional phone-unit HMM. Moreover the prosodic-unit HMM enables us to control the F0 contour more flexibly than phone-unit HMM because it is easy to manipulate the timing of the prosodic events.

4.3 Speech synthesis using prosodic-event-based HMM

In this study, F0 is generated from the prosodic-unit HMM, whereas the other information such as spectral features is generated from phone-unit HMM. However, the positions of phones and prosodic events do not match when the speech parameters are generated independently. Accordingly, the time

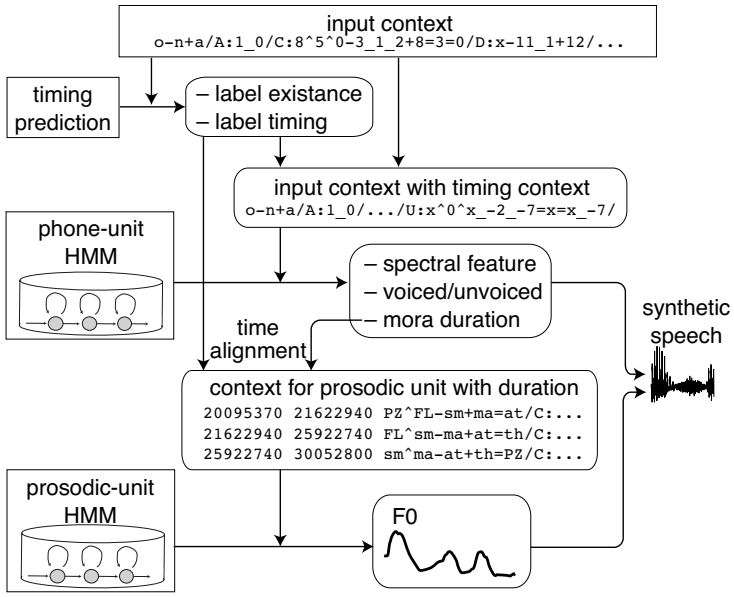


Figure 4.2: An outline of speech synthesis using prosodic-unit HMM.

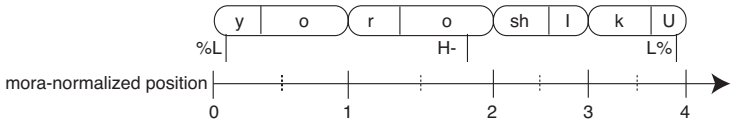


Figure 4.3: An example of mora-normalized position. In this example, the accent phrase consists of 4 moras: “yo”, “ro”, “shi”, and “ku”. The mora-normalized positions of labels, “%L”, “H−”, and “L%” are defined as about 0.1, 1.8, and 3.9, respectively.

alignment between the phone-unit HMM and the prosodic-unit HMM is necessary to apply it to speech synthesis system. Especially, the position of pitch falling by accent is important for Japanese speech synthesis because some words with the same pronunciation are distinguished by it. For this purpose, we use timing prediction of prosodic label and apply it to the alignment between the HMMs. Figure 4.2 shows the procedure of the speech synthesis with the alignment.

In the training step, F0 is modeled by the prosodic-unit HMM. Spectral features, voiced/unvoiced feature, and duration are modeled by the phone-unit HMMs. Here the phone-unit HMMs are trained using the extended contexts proposed in the previous chapter which include the mora positions

of X-JToBI tone tier labels and the tone types, e.g., BPMs, prosodic fillers, and prosodic word fragments. In the timing prediction, the existence of "H−" and "A" and the timing information of each label in the accent phrase are predicted. In this study, we use a mora-normalized position shown in Fig. 4.3 as the label timing information. Mora-normalized position is a measure where the length of each mora is normalized into unity. In Fig. 4.3, the mora-normalized positions of the labels "%L", "H−", and "L%" are defined as about 0.1, 1.8, and 3.9, respectively. As explanatory variables for timing prediction, we use the information of accent phrase.

When synthesizing speech, we firstly construct the input context sequence from the word sequence with automatically or manually annotated prosodic event information such as accent and BPM, and then predict the label timing. Next, the contexts for the phone-unit HMM are constructed using predicted mora-normalized positions, and the spectral, voiced/unvoiced, and duration features are generated from the phone-unit HMMs. Then, the duration information of prosodic units are calculated from the generated mora durations and label timing by the time alignment, and the F0 contour is generated by the prosodic-unit HMMs with their durations. Here, to generate continuous F0 contour, we set a small value as the threshold of the voiced space weight of MSD-HMM. Finally, speech is synthesized using the generated spectral and F0 features.

4.4 Experiments

4.4.1 Experimental conditions

Spontaneous conversational speech data was used for the evaluation experiments. We chose speech data of two female speakers (#19, #514) included in CSJ. Each speaker was non-professional speaker and uttered three sets of conversational speech—two interviews and a task-oriented dialog. The total length of speech samples of each speaker was approximately 25 minutes. Speech signals were sampled at a rate of 16 kHz. The spectral feature and F0 were extracted by STRAIGHT [26] with 5 ms frame shift. The feature vector of prosodic-unit HMM consisted of log F0, and their delta and delta-

delta coefficients. The feature vector of phone-unit HMM consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, their delta and delta-delta coefficients, and a voiced/unvoiced flag. We used hidden semi-Markov model (HSMM) [15] which has explicit duration distributions for both prosodic-unit and phone-unit HMM. The model topology was 5-state left-to-right context-dependent HSMM without skip paths. Each state had a single Gaussian distribution with a diagonal covariance matrix. MDL was used for the stopping criterion. In the case of F0, minimum number of observations was also used to alleviate over-fitting. We set the minimum number of observations to 50 on the basis of a preliminary experimental result. We compared the proposed technique with the conventional HMM-based conversational speech synthesis technique of the previous chapter. In this technique, the phone-unit HMM was used to model both of the spectral and prosodic features.

C4.5 was used for the prediction of the existence of the labels, and linear regression was used for the prediction of mora-normalized position. We chose these classifiers through preliminary experiments. For training and testing, the phonetic and prosodic contexts were automatically converted from the labels given in CSJ. Ten-fold cross-validation tests were performed in the evaluations.

4.4.2 Evaluation of F0 modeling

Performance of the proposed technique was evaluated both objectively and subjectively. To focus on the F0 modeling with the prosodic-unit HMM, F0 patterns were generated using the label timings annotated in the database. Tables 4.3 and 4.5 show the average F0 distortions, correlation coefficients, and tree sizes of log F0 of the proposed and conventional techniques. The average F0 distortion was calculated by RMS error between generated and original log F0s. Table 4.4 shows the result of subjective evaluation of reproducibility. In this test, to focus on the evaluation of F0 reproducibility, we used the acoustic features extracted from the original speech except F0. This test was performed by an XAB test. Six participants chose the sample more similar to the reference X. The reference sample was vocoded speech. When the participants could not determine the preference, “no preference”

Table 4.3: F0 distortion using annotated timing in database.

	RMSE of log F0[cent]	
Speaker	Prosodic-event-based HMM	Phone HMM
#19	288.9	282.9
#514	383.3	381.8

Table 4.4: Subjective evaluation of reproducibility of F0.

Speaker	Prosodic-event-based HMM	No preference	Phone HMM
#19	36.7%	30.0%	33.3%
#514	36.7%	31.7%	31.7%

Table 4.5: The number of leaf nodes of the F0 models.

Speaker	Prosodic-event-based HMM	Phone HMM
#19	262	679
#514	277	762

was chosen. Each participant evaluated 20 utterances randomly chosen from generated speech samples which were used for objective evaluation. We used speech samples having 10 or more moras. It is found from the results that, although average F0 distortions of the proposed technique were larger than those of the conventional technique, the scores of subjective evaluation were comparable. There was no significant difference between the subjective scores of the proposed and conventional techniques. The correlation coefficients were also comparable. It is noted that the number of leaf nodes of the proposed technique, which represents model complexity, was about 36% or 39% as many as the conventional technique. We will give a further discussion about the results in Section 4.5.

4.4.3 Evaluation of synthetic speech

The performance of overall speech synthesis which uses the prosodic-unit HMM and timing prediction was evaluated objectively and subjectively. In

Table 4.6: F0 distortion using predicted timing.

	RMSE of log F0[cent]	
Speaker	Prosodic-event-based HMM	Phone HMM
#19	302.0	288.4
#514	398.2	386.2

Table 4.7: Subjective evaluation of reproducibility of synthetic speech.

Speaker	Prosodic-event-based HMM	No preference	Phone HMM
#19	35.0%	38.3%	26.7%
#514	25.0%	46.7%	28.3%

this experiment, the F0 contour was generated using the technique explained in Section 4.3, i.e., the label timing was predicted and the time alignment between phone-unit and prosodic-unit HMMs was performed. Tables 4.6 and 4.7 show the F0 distortions, correlation coefficients, and the scores of subjective evaluation of reproducibility, respectively. Since the trained HMMs are the same as those of previous subsection, the actual leaf node size is also the same. The conditions of subjective evaluation were the same as the the experiment of Section 4.4.2. It can be found that the results were similar to those of the evaluation of F0 modeling. The scores of reproducibility were comparable between the proposed and conventional techniques and have no significant differences.

4.5 Discussions

As seen in the above results, although the F0 distortions of the proposed technique were larger than those of the conventional technique, the reproducibility by subjective evaluation was comparable. A possible reason is the difference of the characteristics of generated F0, whose example is illustrated in Fig. 4.4. The example includes a Japanese utterance “あのー，う，アンケートを配る時に，大学の先生にたのんだんですよね (anoo, u, aNkeetoo

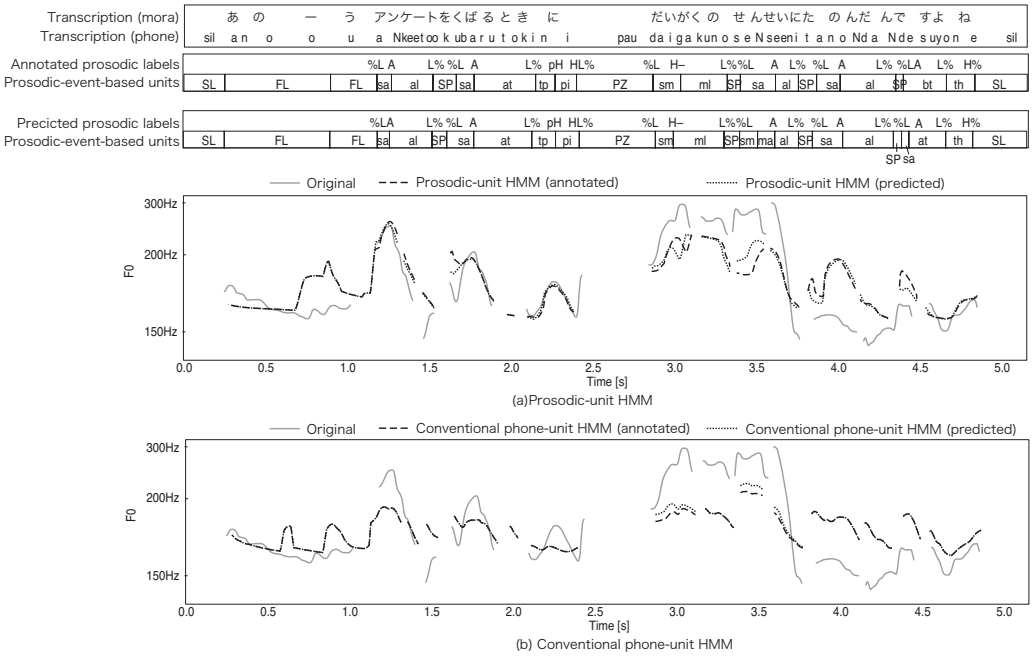


Figure 4.4: Examples of generated F0 contours of the utterance of #514.

kubaru tokini, daigakuno, snseeni, tanoNdaNdesuyone)", which means "So, uh, when distributing the questionnaires, I asked a college teacher." This utterance begins with two fillers "あのー" and "う" followed by two breath groups. The first breath group has a rise-fall BPM "L%–pH–HL%" in the end, and a rise BPM "L%–H%" is observed in the end of the second breath group. Figure 4.4(a) and (b) show the F0 contours generated using the proposed prosodic-unit HMM and the conventional phone-unit HMM, respectively. In the figure, "annotated" and "predicted" in the figure correspond to the cases of Section 4.4.2 and 4.4.3, respectively.

When comparing prosodic-unit HMM with phone-unit HMM, it can be seen in the first breath group that the F0 variation of prosodic-unit HMM is more similar to original one than the flat F0 variation of phone-unit HMM. Such a large F0 variation can be observed in many utterances. As a result, the use of prosodic-unit HMM increased the variance of generated F0 as shown in Table 4.8. This can be attributed to the explicit modeling of the prosodic-unit HMM. However, such variation does not always match with an actual F0 variation. For example, the F0 variation was larger than original

Table 4.8: Standard deviations of generated log F0.

Speaker	Prosodic-unit HMM	Phone-unit HMM	Original
#19	0.119	0.097	0.203
#514	0.142	0.137	0.268

one in "tanoNdaNdesuyone" in Fig. 4.4. This could be a reason why F0 distortion was larger than the conventional method.

The errors of timing prediction would be acceptably small. Indeed, timing prediction enables us to generate similar F0 contours to annotated timing, and it can be found in the figure that the majority of F0 contours were overlapped. Although the mismatches of labels were observed in "senseeni" and "desuyone," the generated F0 contours were similar. This is because the context clustering identifies the contexts that have similar characteristics.

In addition, it can be seen that both prosodic-unit HMM and phone-unit HMM fail to generate a similar F0 contour to original one in the fillers of the begging of the utterance. Since the context of fillers is only a choice of "FL" and "FH", it may be insufficient to model a large variety of fillers. To overcome this problem, it is necessary to incorporate more paralinguistic information, e.g., the function of fillers and intentions. Also, additional information like emphasis is needed for expressing the higher original F0s than the generated ones in "daigakuno senseeni".

4.6 Conclusion

In this chapter, we proposed an F0 modeling technique based on the prosodic-unit HMM. The component of prosodic events was used as a unit of HMM in order to model F0 contour efficiently. The evaluation experiments were performed for both F0 modeling and speech synthesis. The results showed that the subjective reproducibility of the proposed technique was comparable to that of the conventional technique while reducing the leaf node size of F0 model to about 40%.

Chapter 5

Speech Synthesis Based on Gaussian Process Regression

This chapter proposes a statistical speech synthesis technique based on Gaussian process regression (GPR). The model of GPR is designed for directly predicting frame-level acoustic features from corresponding information on frame context that is obtained from linguistic information. The frame context includes the relative position of the current frame and articulatory information and is used as the explanatory variable in GPR. Here, we introduce cluster-based sparse Gaussian processes (GPs), i.e., local GPs and partially independent conditional (PIC) approximation to reduce the computational cost. The experimental results for both isolated phone synthesis and full-sentence continuous speech synthesis revealed that the proposed GPR-based technique without dynamic features significantly outperformed the conventional hidden Markov model (HMM)-based speech synthesis using minimum generation error training with dynamic features.

5.1 Introduction

In corpus-based statistical speech synthesis, parametric speech synthesis based on hidden Markov models (HMMs) [6] has been widely studied [5]. An observation vector sequence consisting of acoustic features is modeled in HMM-based speech synthesis using a hidden state sequence as a generative

model. Although HMM states are discrete-time information, we can generate a smooth and stable speech parameter trajectory by taking dynamic features into account in the model training and speech parameter generation processes [19]. Furthermore, the acoustic characteristics of each speech synthesis unit are represented at the segmental and supra-segmental levels by using context-dependent models where phonetic and prosodic contextual factors are taken into account.

While HMM-based speech synthesis can reflect the acoustic characteristics of training data to synthetic speech using a limited amount of training data, HMM is not always an appropriate model for acoustic features to be synthesized. There are specifically two major problems. First, there is a mismatch where the hidden-state space is discrete despite the continuously changing characteristics of acoustic features. Even though dynamic features enable us to generate a smoothly changing feature trajectory from the discrete states, the parametric representation of acoustic features is limited. In fact, a fixed number of state-dependent dynamic features fail to generate some short-time variations. The second problem is generalization in the model training using decision-tree-based context clustering in which all parameters in each leaf node are tied. Although this improves the estimation accuracy of model parameters and enables model parameters to be predicted for unseen contexts, the resulting number of model parameters is very limited and contextual diversity greatly decreases.

Several techniques have been proposed [33]–[37] to alleviate the quality degradation caused by the above two problems. Rich context modeling [36], [37] is a technique of reducing the over-smoothing effect with the parameter-tying process. The optimum untied HMM sequence for input context labels are searched in this approach by using conventional tied HMMs as guiding models. The subjective quality is expected to be improved when there is a sufficient amount of training data and the contexts of training data adequately cover those of the input texts. In contrast, there are some discontinuities in synthetic speech, which degrade naturalness, when the amount of training data is relatively small because the model parameters are not generalized in model training. Another approach is variance compensation for spectral features using post-filtering [33], [34] or global variance (GV) [35],

and this approach has also been demonstrated to be effective in reducing the buzzy and muffled sounds of synthetic speech. However, spectral distortion between original and synthetic speech generally increases and this occasionally degrades the similarity of synthetic speech.

In recent years, novel approaches using Gaussian processes (GPs) have been proposed to speech processing, such as speech enhancement [38], voice conversion [39], and speech representation [40]. Henter et al. [40] challenged the problem of state discreteness and they extended discrete states to continuous variables of a latent space where GP was used for a frame-level function that transformed the latent space variables into acoustic features. The Gaussian process dynamical model (GPDM) was specifically used to express latent space. However, it is not easy to apply GPDM to text-to-speech directly because of the difficulty of correlating latent space variables with the linguistic information of a given input sentence to be synthesized.

In this chapter, we propose a technique of speech synthesis based on the Gaussian process regression (GPR) [11] to overcome the limitations with parametric models. GPs are known to be nonparametric Bayesian models where "nonparametric" means that model complexity, i.e., the number of parameters, expands with the increase in data size. This implies that GPs are flexible in terms of the complexity of the model. GPs are also robust against over-fitting due to Bayesian inference. In addition, since GPs involve a kernel method, various kinds of data can be used as input variables by defining the kernel function of respective samples [41]. The main advantage of introducing GPR is that we can eliminate parameter tying from model training by directly representing the relation between linguistic and acoustic features using a covariance function of GP. Another advantage is that the acoustic features of each frame can be directly estimated from the frame context that is defined by linguistic information.

Although the proposed technique assumes GP on a frame-level function in the same way as that by Henter et al. [40], there is a difference in that the function transforms frame-level information obtained from linguistic information instead of latent space variables. Here, we define a combined kernel including the kernels for position contexts and for phone contexts for the kernel function of GPs. In addition, we incorporate approximation techniques

into GPs to achieve feasible computation of GP training.

5.2 Speech synthesis based on Gaussian process regression for isolated phones

This section briefly describes the basic theory of general GPR [11] and then presents the framework of GPR-based speech synthesis for a small amount of speech data, i.e., isolated phone segments. A frame context kernel is designed as an input variable of the GPR to represent frame-level acoustic features. This framework is then extended to full-sentence continuous speech synthesis, which is discussed in Section 5.4.

5.2.1 Gaussian process for regression

Suppose that we have a training data set, $\mathcal{D} = \{(\mathbf{x}_n, y_n)|n = 1, \ldots, N\}$, and a test data set, $\mathcal{D}_T = \{(\mathbf{x}_t, y_t)|t = 1, \ldots, T\}$, where $\mathbf{x}_n$ is a column vector consisting of explanatory (input) variables, and y_n is an output scalar variable. We assume that y_n is given by

$$y_n = f(\mathbf{x}_n) + \epsilon \tag{5.1}$$

where $f(\mathbf{x}_n)$ is a noise-free latent function value and ϵ represents the Gaussian noise of $\mathcal{N}(0, \sigma^2)$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$ and $\mathbf{y} = [y_1, \ldots, y_N]^\top$ be matrix forms of all input and output variables of training data and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^\top$ be the latent function values of the training data. We define $\mathbf{X}_T$, $\mathbf{y}_T$, and $\mathbf{f}_T$ as matrix forms for test data in the same way as the training data.

When $f(\mathbf{x}_i)$ is a Gaussian process, the GP prior is given by

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_N + \sigma^2\mathbf{I}) \tag{5.2}$$

where $\mathbf{K}_N$ is a Gram (or covariance) matrix of the training data whose element is given by

$$K_{mn} = k(\mathbf{x}_m, \mathbf{x}_n) \quad m = 1 \ldots N,\ n = 1 \ldots N \tag{5.3}$$

and $k(\mathbf{x}_m, \mathbf{x}_n)$ is a kernel (or covariance) function.

The main goal of GPR is to infer the continuous distributions of output variables of test data, $\mathbf{y}_T$, given new input vectors $\mathbf{X}_T$. The joint distribution on the function values, $\mathbf{f}$ and $\mathbf{f}_T$, of the training and test data is given by

$$p(\mathbf{f}, \mathbf{f}_T|\mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{N+T}), \tag{5.4}$$

$$\mathbf{K}_{N+T} = \begin{bmatrix} \mathbf{K}_N & \mathbf{K}_{NT} \\ \mathbf{K}_{TN} & \mathbf{K}_T \end{bmatrix} \tag{5.5}$$

where $\mathbf{K}_T$ is a Gram matrix of test frames, and Gram matrix $\mathbf{K}_{NT} = \mathbf{K}_{TN}^\top$ consists of covariances between the training and test frames.

The joint distribution of $\mathbf{y}$ and $\mathbf{y}_T$ is given by

$$p(\mathbf{y}, \mathbf{y}_T|\mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{N+T} + \sigma^2\mathbf{I}). \tag{5.6}$$

Given a training data set, the predictive distribution of output variables of a test data set is obtained by

$$p(\mathbf{y}_T|\mathbf{y}, \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T) \tag{5.7}$$

$$\boldsymbol{\mu}_T = \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \tag{5.8}$$

$$\boldsymbol{\Sigma}_T = \mathbf{K}_T - \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{K}_{NT}. \tag{5.9}$$

The inversion of $(\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}$ requires $\mathcal{O}(N^3)$ computations. For practical implementation, the parameter vector

$$\boldsymbol{\alpha} = [\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \tag{5.10}$$

that only depends on the training data set is calculated in the training step. The number of parameters in $\boldsymbol{\alpha}$ is N, which corresponds to the number of frames of the training data. From (5.8), a set of new output means is given by an inner product

$$\boldsymbol{\mu}_T = \mathbf{K}_{TN}\boldsymbol{\alpha} \tag{5.11}$$

which requires $\mathcal{O}(N)$ computational cost.

We need to design the kernel function to use GP for regression. The necessary conditions for the kernel function are that the Gram matrix be positive

semi-definite and symmetric. We used two typical kernels, i.e., square exponential (SE) kernel and linear kernel in this research. The SE kernel is the most widely used stationary kernel as the measure of "similarity" between two input vectors. The SE kernel is defined by

$$k(\mathbf{x}_m, \mathbf{x}_n) = \exp\left(-\frac{\|\mathbf{x}_m - \mathbf{x}_n\|^2}{l^2}\right) \tag{5.12}$$

where l denotes a length-scale hyper-parameter. The linear kernel is given by

$$k(\mathbf{x}_m, \mathbf{x}_n) = \mathbf{x}_m^\top \mathbf{x}_n. \tag{5.13}$$

This kernel assumes linearity between output and input features. It should also be noted that it is possible to construct a new kernel by combining multiple arbitrary kernel functions by means of some operations such as sum, product, and convolution [41].

5.2.2 Frame context with kernel design

We use frame-level features obtained from the linguistic information of transcriptions for the explanatory variables of the regression model. We define *frame context* that includes the relative position and phonetic information of the current frame as

$$\mathbf{x}_n = (p_n, \mathbf{c}_n). \tag{5.14}$$

Fig. 5.1 has an example of the frame context. A normalized relative position in the current phone is employed for position context p_n, where the beginning of the phone is set to zero and its end is set to one. We use a set of preceding, current, and succeeding phonetic features for phone context $\mathbf{c}_n$. We introduce binary variables ($\{\text{positive} = +1, \text{negative} = -1\}$) for each phonetic feature listed in Table 5.1 based on a balanced distinctive phonetic feature set [42]. Let P be the number of phonetic features; then, a $3P$-dimensional binary-valued vector is constructed.

The proposed frame context kernel is defined as a product of two kernels.

$$k(\mathbf{x}_m, \mathbf{x}_n) = k_p(p_m, p_n) k_c(\mathbf{c}_m, \mathbf{c}_n) \tag{5.15}$$

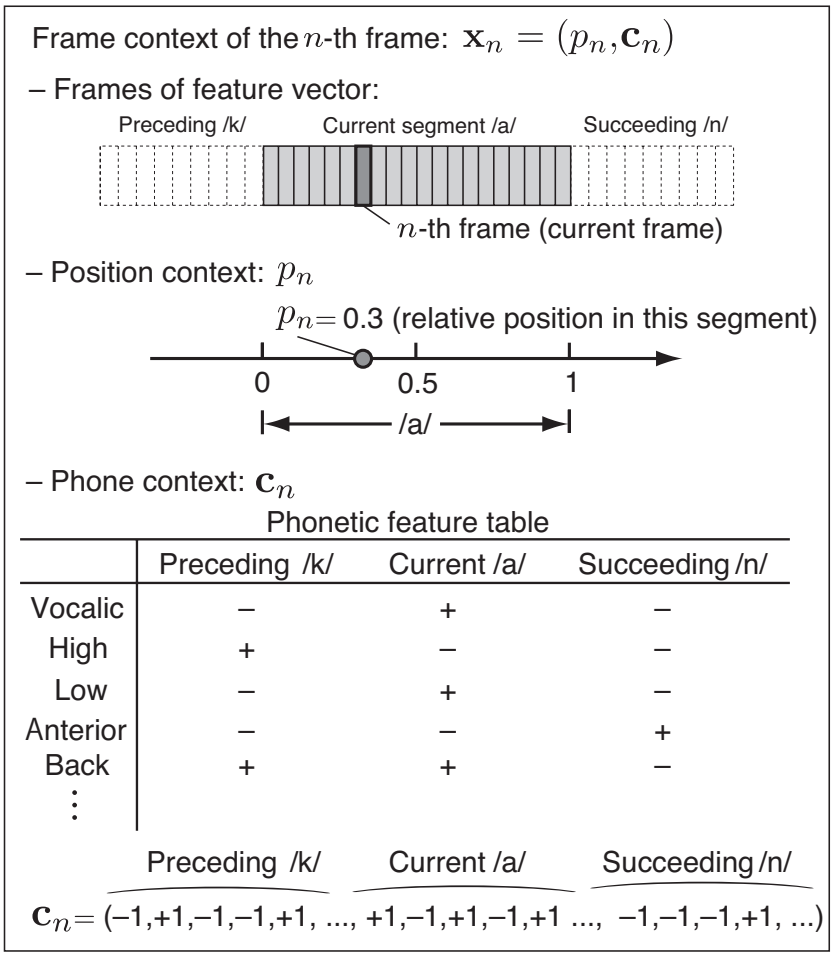


Figure 5.1: Example of frame context, i.e., frame-level input variable set for GP regression. This example has frame context for frame positioned in phone /a/, which is between preceding phone /k/ and succeeding phone /n/.

Table 5.1: Binary phonetic features used for phone context kernel.

Phonetic features
vocalic, high, low, anterior, back, coronal, plosive, affricative, continuant, voiced, nasal, semi-vowel, silent

where $k_p(p_m, p_n)$ and $k_c(\mathbf{c}_m, \mathbf{c}_n)$ correspond to the position kernel and the phone context kernel. The position kernel represents the similarity of position contexts in phones whereas the phone context kernel represents that of phone

contexts.

5.2.2.1 Position kernel

The SE kernel is used for the position kernel and is given by

$$k_p(p_m, p_n) = \exp\left(-\frac{(p_m - p_n)^2}{l_p^2}\right), \tag{5.16}$$

where p_m is the relative position of the m-th frame.

5.2.2.2 Phone context kernel

We examine two different phone context kernels in this chapter. The first is the sum of SE kernels and the second is a linear kernel. The former one is defined by

$$k_c(\mathbf{c}_m, \mathbf{c}_n) = \sum_{k=1}^{3P} \theta_{ck}^2 \exp\left(-\frac{(c_{mk} - c_{nk})^2}{l_{ck}^2}\right) \tag{5.17}$$

where l_{ck} is a scale hyper-parameter, and θ_{ck} is a hyper-parameter that represents the relevance of the k-th phonetic feature. The kernel value becomes maximum when the input phone contexts are the same.

The linear kernel is given by

$$k_c(\mathbf{c}_m, \mathbf{c}_n) = \sum_{k=1}^{3P} \theta_{ck}^2 c_{mk} c_{nk}. \tag{5.18}$$

The use of this kernel assumes that acoustic features in the same position are on a hyperplane in the $3P$-dimensional phonetic feature space.

5.2.3 GPR-based speech synthesis

Fig. 5.2 outlines a basic GPR-based speech synthesis system. While an acoustic feature is generally a multi-dimensional vector, here we have assumed that all dimensions are independent and each dimension can be modeled separately. When synthesizing speech, we generate a single feature sequence from the predictive distribution with a certain method, such as using the mean sequence for synthetic parameters or generating random sequences from the

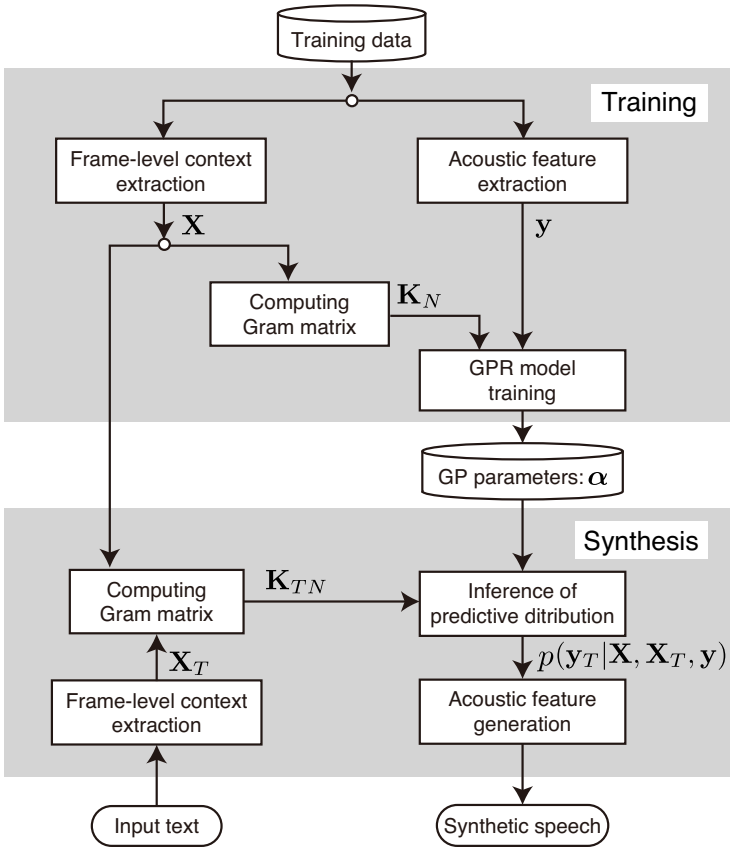


Figure 5.2: Outline of speech synthesis process in proposed approach.

distribution. We adopt the mean sequence, $\boldsymbol{\mu}_T$, of the distribution in this study.

Consequently, the training and synthesis procedures are summarized as:

Training step

1. Frame-level acoustic features such as mel-cepstral coefficients and fundamental frequency are extracted from the training data.

2. The frame contexts are created from the annotation data including the phone boundaries of the training data.

3. Gram matrix $\mathbf{K}_N$ between the frames of the training data is determined using the frame contexts.

4. Parameter vector $\boldsymbol{\alpha}$ in (5.10) is calculated using $\mathbf{K}_N$.

Synthesis step

1. The frame contexts are created from the input sentence.
2. Gram matrix $\mathbf{K}_{TN}$ between the frames of the training and new input data is calculated.
3. The mean sequence $\boldsymbol{\mu}_T$ of the predictive distribution is calculated by multiplying Gram matrix $\mathbf{K}_{TN}$ and $\boldsymbol{\alpha}$ and is used as a generated spectral feature trajectory.
4. The output waveform is synthesized using the spectral and excitation features.

5.3 Experiments on isolated phone synthesis

5.3.1 Experimental conditions

The speech database used in the experiments consisted of 503 ATR phonetically balanced Japanese sentences recorded by one female speaker. The speech signals were sampled at a rate of 16 kHz. The spectral features were extracted with STRAIGHT [26]. The 0-39th mel-cepstral coefficients were used as output variables. Each dimension of the mel-cepstral coefficients was modeled separately.

We chose five vowels (/a/, /i/, /u/, /e/, and /o/) and five consonants (/k/, /s/, /t/, /n/, and /m/), which are primary phonemes in Japanese, to examine the potential of GPR. Each phone was segmented using manually annotated phone boundaries. The phone segments of the training set were randomly chosen up to 10,000 frames from 450 sentences for each phoneme. The 50 test phone segments were randomly chosen from the remaining 53 sentences. The manually annotated boundaries of the original utterances were given when the test segments were synthesized.

The following experiments were performed separately for each phoneme. The HMM-based speech synthesis was used as a conventional technique. Triphones were used for the context set for HMM training. The model topology was a five-state, left-to-right, no-skip hidden semi-Markov model (HSMM).

Table 5.2: Average spectral distortions of generated parameter sequences using position context for primary phonemes. Values represent mel-cepstral distances [dB].

	monophone			monophone	
Phoneme	HMM	GPR	Phoneme	HMM	GPR
a	6.02	6.08	k	6.02	5.98
i	7.11	7.09	t	4.35	4.41
u	7.18	7.16	n	6.27	6.28
e	6.04	6.07	s	5.18	5.03
o	6.48	6.48	m	5.92	5.94

Each state had a single Gaussian distribution with a diagonal covariance matrix and the feature vector included delta and delta-delta dynamic features.

5.3.2 Evaluation of position kernel

An objective evaluation was done under the condition where only the position context was given as the input to assess the performance of GPR in generating continuously changing acoustic features. The kernel for GPR corresponds to that given in (5.16). All acoustic features were normalized by their means and variances. The hyper-parameter l_p was set to 0.289, which was the standard deviation of frame contexts in the training data, and noise parameter σ was set to 1.0 according to preliminary results of the experiments.

The spectral distortions of the generated parameter sequences from both techniques are summarized in Table 5.2. In the table, GPR in the table represents the proposed GP regression. The average mel-cepstral distance was used as the measure of spectral distortion. From the table, It can be seen from the table that GPR perform es comparably to HMM. This means that GPR generated continuously changing acoustic features even though dynamic features were not used for this regression.

Table 5.3: Average spectral distortions of generated parameter sequences using frame context. Values represent mel-cepstrum distances [dB].

Phoneme	triphone HMM	GPR-SE	GPR-linear
a	5.67	5.51	5.52
i	6.01	5.64	5.63
u	6.10	5.94	5.94
e	5.33	5.17	5.16
o	5.90	5.63	5.64
k	5.09	5.05	5.05
t	4.13	4.17	4.17
n	5.73	5.81	5.81
s	4.74	4.57	4.57
m	5.48	5.50	5.50

5.3.3 Evaluation of frame context kernel

We then evaluated the proposed frame context kernels described in Section 5.2.2. We compared the sum of SE kernels and the linear kernel as the phone context kernel. All output variables were normalized and the hyper-parameters were given by $l_p = l_{ck} = 0.289$ $(k = 1, \ldots, 3P)$, $\sigma = 1.0$, and $\theta_{ck} = 1.0/3P$ $(k = 1, \ldots, 3P)$ on the basis of the preliminary experimental results. Triphone HMM was used for HMM and decision-tree-based context clustering was carried out with an MDL criterion [18].

Table 5.3 lists the mel-cepstral distances between the generated and original sequences.

GPR-SE and GPR-linear employed the sum of SE kernels for the former and the linear kernel for the latter for the phone context kernel. It could be confirmed that phone context reduced the distortions with all techniques compared to the results in Table 5.2 where the phonetic context was not used. Even though there were only small differences for consonants except /s/ in comparing GPR with HMM, the mel-cepstral distances for the vowels using GPR-SE and GPR-linear significantly decreased. We also found that the distances for GPR-SE and GPR-linear were comparable. One possible reason is that the characteristics of kernel values were similar between the

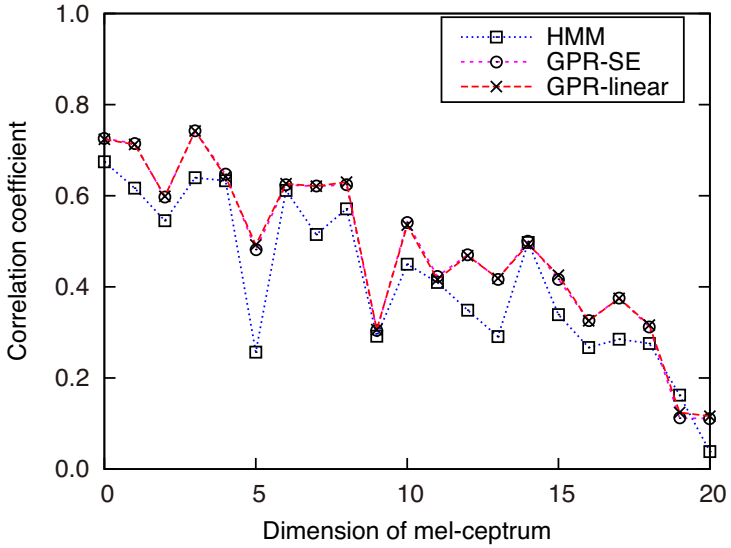


Figure 5.3: Correlation coefficients between generated and original mel-cepstral coefficients for phoneme /i/.

sum of SE kernels and the linear kernel under the condition that the binary-valued vectors were used as input variables.

The correlation coefficients between generated and original acoustic features for each mel-cepstral dimension are plotted in Fig. 5.3 to enable a more detailed look. The 0-20th dimensions are given in the figure, because the correlation coefficients of higher dimensions were too low to discuss. The results for GPR-SE and GPR-linear are very similar and they almost overlap. We can see that the correlations for GPR-SE and GPR-linear are higher than those for HMM in most dimensions.

5.4 Continuous speech synthesis based on sparse Gaussian processes

The matrix inversion needs $\mathcal{O}(N^3)$ calculations in the training procedure to obtain parameter $\boldsymbol{\alpha}$ in (5.10). The value of N is generally at least hundreds of thousands[1]. Therefore the computational complexity of GPR for continuous

[1]If we have 10 min of speech data with 5-ms shift, N is 120,000.

speech synthesis is not realistic. We examine two methods of approximation to reduce the computational cost, i.e., local GPs [43], [44] and partially independent conditional (PIC) approximation [44]. These methods enable feasible computation by approximating matrices to be sparse. While there are various kinds of approximation methods, e.g., subset of data (SoD) [11], [45] and fully independent training conditional (FITC) approximation [45], we chose the local GPs and PIC because they are effective methods of modeling local characteristics within phone segments.

5.4.1 Local GPs

The use of Local GPs involves a method of reducing the amount of computation by simply dividing all the data into local blocks and modeling each block separately. That is, covariance matrix $\mathbf{K}_{N+T}$ is approximated by block diagonal one:

$$\mathbf{K}_{N+T} \approx \mathbf{K}_{N+T}^{\mathrm{LOCAL}} = \mathrm{diag}\left[\mathbf{K}_{B_1}, \mathbf{K}_{B_2}, \ldots, \mathbf{K}_{B_S}\right]. \tag{5.19}$$

When all training frames are divided into S blocks and each block has at most B training frames, the computational cost results in $\mathcal{O}(SB^3)$. By fixing B, the computational complexity increases linearly with the number of training data N.

In order to use the local GPs, it is necessary not only to determine the block of the training frames but also that of the synthesis frames from their linguistic features. We utilize decision-tree-based context clustering in this study, which is effectively used in HMM-based speech modeling. We stop splitting nodes if a node has less than B frames when constructing the decision tree. We conduct phone-level clustering instead of state-level or stream-level clustering because the state and stream information is unknown in the synthesis step.

The local GPs and the HMM-based speech synthesis both use the decision tree clustering of context dependent HMMs. In the HMM-based technique, the observation samples of each cluster are collected and converted into a limited number of distributions. In contrast, in GPR with local GPs, the covariances of the samples in the sames cluster are calculated and training each cluster with GPR yields at most B parameters.

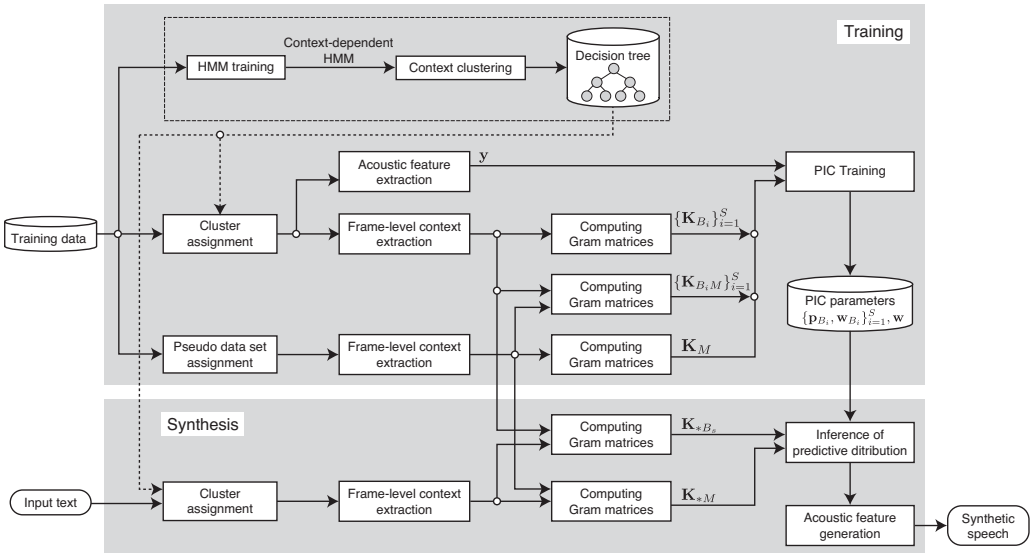


Figure 5.4: Overview of training and synthesis stages in GPR-based speech synthesis using PIC approximation.

5.4.2 Partially independent conditional (PIC) approximation

Although the local GPs can model internally changing features effectively within the blocks, the covariances between different blocks are completely ignored. On the other hand, a partially independent conditional (PIC) approximation estimates the covariances between different blocks using a *pseudo-data set*. Pseudo-data set $\bar{\mathcal{D}} = \{(\bar{\mathbf{x}}_m, \bar{y}_m) | m = 1, \ldots, M\}$ is a small amount of data set with a size of $M \ll N$, and the pseudo-data are expected to be distributed similarly to the training data. PIC is a kind of the approximation methods called the sparse pseudo-input Gaussian process (SPGP) [46]. The joint distribution of the function values, $\mathbf{f}$ and $\mathbf{f}_T$, is given by a marginal distribution for pseudo-data variables $\bar{\mathbf{f}} = [f(\bar{\mathbf{x}}_1), \ldots, f(\bar{\mathbf{x}}_N)]^\top$ as

$$p(\mathbf{f}, \mathbf{f}_T) = \int p(\mathbf{f}, \mathbf{f}_T | \bar{\mathbf{f}}) p(\bar{\mathbf{f}}) d\bar{\mathbf{f}} \tag{5.20}$$

where both $p(\mathbf{f}, \mathbf{f}_T|\bar{\mathbf{f}})$ and $p(\bar{\mathbf{f}})$ follow Gaussian distributions and are given by

$$p(\mathbf{f}, \mathbf{f}_T|\bar{\mathbf{f}}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \tag{5.21}$$

$$\bar{\boldsymbol{\mu}} = \mathbf{K}_{(N+T)M}\mathbf{K}_M^{-1}\bar{\mathbf{f}} \tag{5.22}$$

$$\bar{\boldsymbol{\Sigma}} = \mathbf{K}_{N+T} - \mathbf{K}_{(N+T)M}\mathbf{K}_M^{-1}\mathbf{K}_{M(N+T)} \tag{5.23}$$

$$p(\bar{\mathbf{f}}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_M) \tag{5.24}$$

where $\mathbf{K}_{(N+T)M}$ is a Gram matrix between the frames of all data $(\mathbf{X}, \mathbf{X}_T)$ and the pseudo-data, and $\mathbf{K}_M$ is a self covariance matrix of the pseudo-data set. SPGP is a method of avoiding the direct calculation of matrix inversion in (5.11) by approximating $p(\mathbf{f}, \mathbf{f}_T|\bar{\mathbf{f}})$. The $\bar{\boldsymbol{\Sigma}}$ is approximated in PIC by using a block diagonal matrix as

$$\bar{\boldsymbol{\Sigma}} \approx \bar{\boldsymbol{\Sigma}}^{\mathrm{PIC}} = \mathrm{diag}\left[\boldsymbol{\Sigma}_{B_1}, \boldsymbol{\Sigma}_{B_2}, \ldots, \boldsymbol{\Sigma}_{B_S}\right]. \tag{5.25}$$

The Gram matrix of training data is approximated by

$$\mathbf{K}_N \approx \mathbf{K}_N^{\mathrm{PIC}} = \begin{bmatrix} \mathbf{K}_{B_1} & \mathbf{Q}_{B_1B_2} & \cdots & \mathbf{Q}_{B_1B_S} \\ \mathbf{Q}_{B_2B_1} & \mathbf{K}_{B_2} & & \mathbf{Q}_{B_2B_S} \\ \vdots & & \ddots & \vdots \\ \mathbf{Q}_{B_SB_1} & \mathbf{Q}_{B_SB_2} & \cdots & \mathbf{K}_{B_S} \end{bmatrix} \tag{5.26}$$

where $\mathbf{Q}_{B_iB_j}$ is a matrix given by

$$\mathbf{Q}_{B_iB_j} = \mathbf{K}_{B_iM}\mathbf{K}_M^{-1}\mathbf{K}_{MB_j}. \tag{5.27}$$

The $\mathbf{K}_{B_iM}$ and $\mathbf{K}_{MB_j}$ are Gram matrices whose elements are kernel values between the samples of the clustered block and the pseudo-data set. Specifically, the approximation avoids direct calculations of inter-block Gram matrices by means of the pseudo-data set.

When a new input value, $\mathbf{x}_*$, for a certain frame is assigned to cluster B_s, the corresponding mean for $\mathbf{x}_*$ is given by

$$\mu_* = \mathbf{K}_{*M}(\mathbf{w} - \mathbf{w}_{B_s}) + \mathbf{K}_{*B_s}\mathbf{p}_{B_s} \tag{5.28}$$

where $\mathbf{K}_{*M}$ is a Gram matrix between the frames of $\mathbf{x}_*$ and the pseudo-data set, and $\mathbf{K}_{*B_s}$ is a Gram matrix between the frames of $\mathbf{x}_*$ and the s-th block

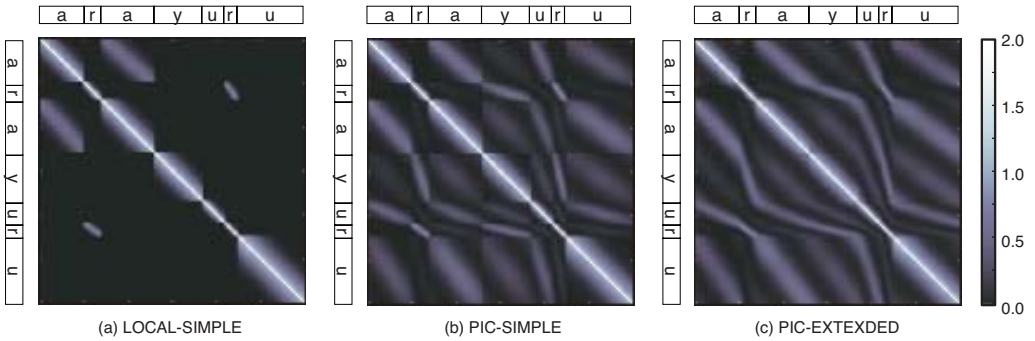


Figure 5.5: Example of covariance matrices of Japanese phrase segment "a r a y u r u" using (a) local GPs and simple frame context, (b) PIC and simple frame context, and (c) PIC and extended frame contexts.

data. The first and second terms on the right-hand side of (5.28) correspond to global and local acoustic characteristics. The $\mathbf{w}$, $\mathbf{w}_{B_s}$ and $\mathbf{p}_{B_s}$ are PIC model parameters calculated by

$$\mathbf{w} = \sum_{s=1}^{S} \mathbf{w}_{B_s} \tag{5.29}$$

$$\mathbf{w}_{B_s} = \mathbf{K}_M^{-1} \mathbf{K}_{MB_s} \mathbf{p}_{B_s} \tag{5.30}$$

$$[\mathbf{p}_{B_1}^\top \cdots \mathbf{p}_{B_S}^\top]^\top = [\mathbf{K}_N^{\mathrm{PIC}} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y}. \tag{5.31}$$

When the maximum block size is B, the number of blocks is S, and the number of frames of the pseudo-data set is M, the computational cost results in $\mathcal{O}(S(B^3 + M^3))$. Methods of determining the blocks and the pseudo-data set are needed to use PIC. The blocks of frames are determined in the same way as the local GPs. We adopt random selection from the training data to select the pseudo-data set.

There is an overview of speech synthesis using PIC approximation in Fig. 5.4. In the training procedure, first, the decision tree of contexts is constructed using context-dependent HMMs. Then the pseudo-data set is chosen from the training data, and the cluster for each training data frame is assigned by the decision tree. Gram matrices are computed after that. PIC parameters $\{\mathbf{p}_i\}_{i=1}^S$, $\{\mathbf{w}_i\}_{i=1}^S$, and $\mathbf{w}$ in (5.29) to (5.31) are calculated at the end of the training procedure. When speech is synthesized, the cluster for each frame context extracted from an input text is also determined by the

decision tree. Next, Gram matrices between synthesis and training frames are computed and the acoustic features of the frames are generated from the Gram matrices and trained PIC parameters. Finally, a speech utterance is synthesized by using the spectral feature trajectory that is generated.

5.4.3 Extension of frame context using adjacent phones

Even though PIC can express the covariances between different blocks, the simple frame context proposed in Section 5.2.2 is insufficient for synthesizing natural-sounding speech. A problem occurs when the simple frame context is used where covariances at the boundary of adjacent phones become discontinuous. For example, the context of the first frame of a current phone and that of the last frame of the preceding phone are entirely different. The discontinuity in covariance causes synthetic speech to be rough.

We extend the frame context to include smoothly changing values in order to overcome the discontinuity in covariance. Since a certain frame not only has information on the current phone but also that on nearby phones, extended frame context $\mathbf{x}_n$ is defined as a set of position and phone contexts of adjacent phones.

$$\mathbf{x}_n = (\mathbf{w}_n, \mathbf{p}_n, \mathbf{C}_n) \tag{5.32}$$

where $\mathbf{w}$, $\mathbf{p}$, and $\mathbf{C}$ are sets of weights, position contexts, and phone contexts expressed as

$$\mathbf{w}_n = \{w_n^{(-1)}, w_n^{(0)}, w_n^{(+1)}\} \tag{5.33}$$

$$\mathbf{p}_n = \{p_n^{(-1)}, p_n^{(0)}, p_n^{(+1)}\} \tag{5.34}$$

$$\mathbf{C}_n = \{\mathbf{c}_n^{(-1)}, \mathbf{c}_n^{(0)}, \mathbf{c}_n^{(+1)}\}. \tag{5.35}$$

The superscripts -1, 0, and $+1$ of the variables correspond to the preceding, current, and succeeding phones. Note that $p_n^{(0)}$ and $\mathbf{c}_n^{(0)}$ correspond to p_n and $\mathbf{c}_n$ in Section 5.2.2. The $p_n^{(-1)}$, $p_n^{(0)}$, and $p_n^{(+1)}$ respectively represent the normalized relative positions of the current frame in the preceding, current, and succeeding phones. The $p_n^{(-1)}$ equals $p_n^{(0)}+1$ and $p_n^{(+1)}$ equals $p_n^{(0)}-1$. The

$\mathbf{c}_n^{(-1)}$, $\mathbf{c}_n^{(0)}$, and $\mathbf{c}_n^{(+1)}$ correspond to the phone context of preceding, current, and succeeding phones. The $w_n^{(i)}$ represents the weight used to emphasize the effect of closer phones. The following sine window function is used in this study as a weight to limit the effect of position and phone kernels to that of nearby phones with smoothly changing weight values.

$$w_n^{(i)} = \begin{cases} \sin\left(\pi(p_n^{(i)} + 0.5)/2\right) & -0.5 \leq p_n^{(i)} \leq 1.5 \\ 0 & \text{otherwise.} \end{cases} \tag{5.36}$$

We use a convolution kernel [47], which computes the sum of all combinations between the adjacent phones of two input variables. The kernel function for the extended contexts is given by

$$k(\mathbf{x}_m, \mathbf{x}_n) = \sum_{i\in\{-1,0,+1\}} \sum_{j\in\{-1,0,+1\}} \left[w_m^{(i)} w_n^{(j)} k_p(p_m^{(i)}, p_n^{(j)}) k_c(\mathbf{c}_m^{(i)}, \mathbf{c}_n^{(j)}) \right]. \tag{5.37}$$

Fig. 5.5 shows examples of covariance matrices. Since the local GPs are used in Fig. 5.5 (a), many of the elements are zero because only intra-cluster covariances are defined. Inter-cluster covariances are estimated by using PIC in Fig. 5.5 (b) and (c). Moreover, the extended context in Fig. 5.5 (c) yields smooth covariances around the boundary of adjacent phones.

5.5 Experiments on continuous speech synthesis

5.5.1 Experimental conditions

The speech database used in the experiments for continuous speech synthesis was the same as that used in Section 5.3. Speech signals were sampled at a rate of 16 kHz, and the frame shift was 5 ms. Spectral envelope, F0, and aperiodicity features were extracted by using STRAIGHT [26]. The 0-39th mel-cepstral coefficients were used as output variables, and each dimension of the mel-cepstral coefficients was modeled separately, which was also the same condition as that in Section 5.3. Speech samples were synthesized using generated spectral features and original F0 and phone durations.

The maximum number of frames, B, of the cluster described in Section 5.4.2 was set to 1000 for the local GPs and PIC, and the size of pseudo-data set M was set to 200. The linear kernel was used for the position kernel. The hyper-parameters were set to $l_p = 0.289$, $\sigma = 1.0$, and $\theta_{ci} = 1.0/3P$ $(i = 1, \ldots, 3P)$, which were the same settings as those in Section 5.2.2. The $\mathbf{K}_M$ must be positive definite to calculate the $\mathbf{K}_M^{-1}$ of (5.30) in PIC, and hence the value of $\theta_\delta \delta_{mn}$ was added to the kernel function $k(\mathbf{x}_m, \mathbf{x}_n)$ where δ_{mn} was the Kronecker delta. The θ_δ was set to unity on the basis of preliminary experimental results.

We also evaluated HMM-based speech synthesis with and without minimum generation error (MGE) training [48] for comparison. The model topology and feature vector including dynamic features were the same as those in Section 5.3. Triphones were used for the context set for HMM training and the MDL criterion was used for context clustering.

5.5.2 Objective evaluation

First, we objectively compared the performance of the conventional and proposed techniques. The mel-cepstral distance between synthetic and original speech were used as an objective distortion measure. We used 150, 250, 350, and 450 sentences as the training data, and 53 sentences were used as the test data. The test data were not included in the training data. We compared two kinds of HMM-based techniques and three kinds of proposed GPR-based techniques. The results are plotted in Fig. 5.6. The "HMM" in the figure represents the HMM-based technique where the model parameters were optimized by the maximum likelihood (ML) criterion. The "HMM-MGE" used MGE training to optimize the model parameters. In the proposed GPR-based techniques, L and P corresponded to local GPs and PIC for approximation, and S and E denote the simple frame context and the extended frame context. We can see that both HMM-MGE and GPR-based techniques yielded smaller distortions than HMM. Moreover, the GPR-based techniques derived significantly smaller distortions than HMM-MGE, which means that frame-level regression performed well. We could see that distortions decreased slightly for all the training sets by comparing GPR-LS and

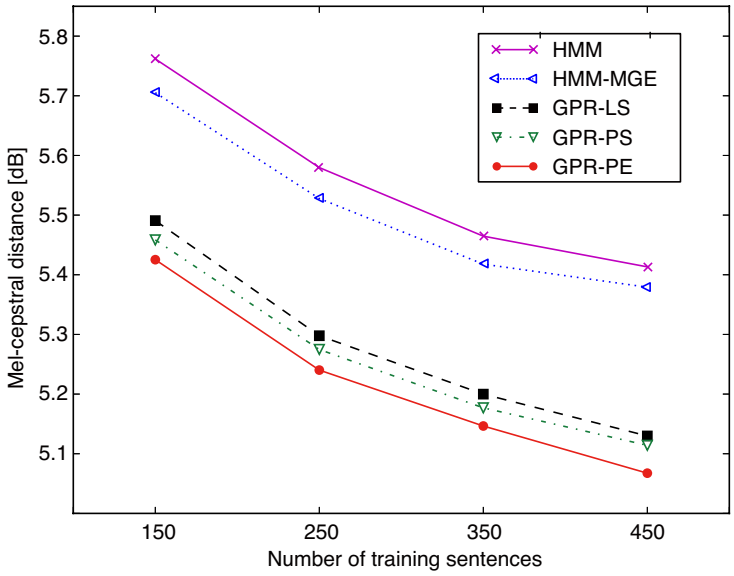


Figure 5.6: Average spectral distortions between original and synthetic speech as function of the number of training sentences.

GPR-PS. In addition, GPR-PE had consistently higher reproducibility than GPR-PS.

Next, we compared the HMM-MGE and proposed GPR-PE techniques in terms of the correlation between the original and generated mel-cepstral coefficients. There were 450 training sentences, and correlation coefficients were calculated for each dimension. Fig. 5.7 plots the results. We can see that the correlation coefficients of the proposed GPR-based technique were higher than those of the HMM-based technique in the most dimensions when we compare them. One reason for this improvement is that the GPR-based technique can directly infer the features of respective frames whereas the HMM-based technique cannot.

5.5.3 Subjective evaluation

5.5.3.1 Naturalness

We evaluated HMM-MGE, GPR-LS, and GPR-PE by using a mean opinion score (MOS) test to subjectively examine the naturalness of the synthetic speech samples. There were 450 training sentences. Five participants lis-

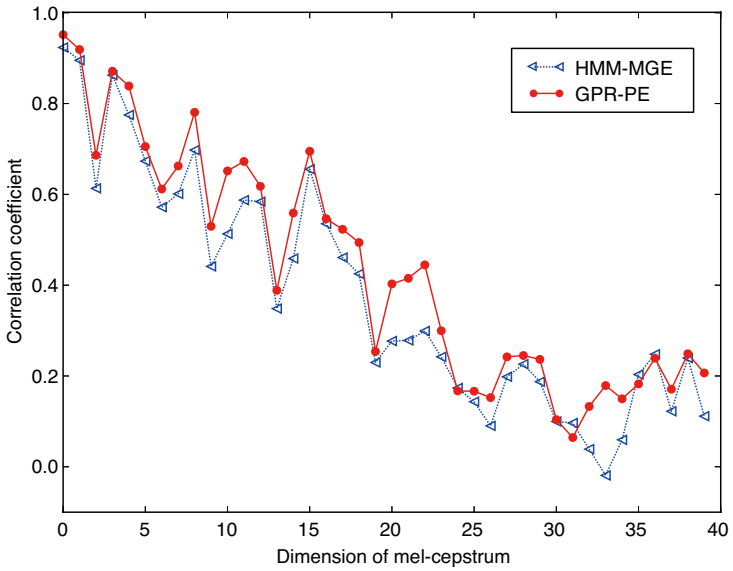


Figure 5.7: Correlation coefficients between original and generated mel-cepstral coefficients.

tened to the synthetic speech samples and rated the naturalness of synthetic speech on a five-point scale, i.e., 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. Ten sentences were randomly chosen from the 53 sentences for each participant. Fig. 5.8 has the mean opinion score (MOS). The error bars indicate 95% confidence intervals. We can see that the scores were comparable when comparing GPR-LS and HMM-MGE, whereas GPR-LS derived smaller mel-cepstral distances in the objective evaluation. This is because the generated acoustic features were not smooth at the phone boundaries and this discontinuity degraded naturalness. In contrast, GPR-PE, which provided continuity on covariance matrices, yielded the highest score for the three techniques.

5.5.3.2 Similarity

We conducted an XAB test on speech similarity between vocoded and synthetic speech samples to compare the reproducibility of synthetic speech samples with the conventional and proposed techniques. We compared HMM-MGE, GPR-LS, and GPR-PE that were evaluated in Section 5.5.3.1. The

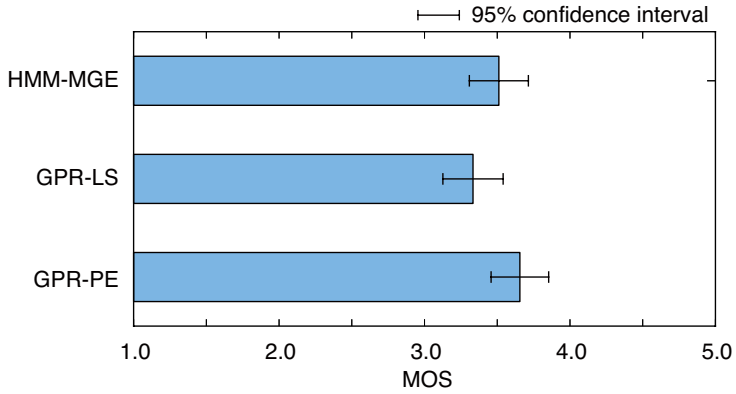


Figure 5.8: MOS on naturalness of synthetic speech.

participants were five Japanese native speakers, and ten sentences were randomly chosen from the 53 test sentences for each of the participants. After being given a vocoded speech sample (X) as a reference, the participants listened to two synthetic speech samples (A and B) in random order and were asked whether A or B was closer to X. We used synthetic speech samples with all three combinations of the three techniques for the pairs of A and B. Fig. 5.9 has the average preference scores for each technique with confidence intervals of 95%. There are no significant differences in the scores of the three techniques in the figure. However, the average scores for GPR-PE and GPR-LS are about 0.1 point higher than the score for HMM-MGE, which indicates that the proposed GPR-based speech synthesis is comparable or slightly better than the conventional HMM-based speech synthesis in terms of reproducibility.

5.6 Conclusion

This chapter proposed a novel approach to speech synthesis using Gaussian process regression (GPR). We first described the basic framework of GPR-based speech synthesis and evaluated it using a small data set of isolated phones. We then achieved continuous speech synthesis with feasible computational cost using partially independent conditional (PIC) approximation and context extension. The evaluation results revealed that introducing

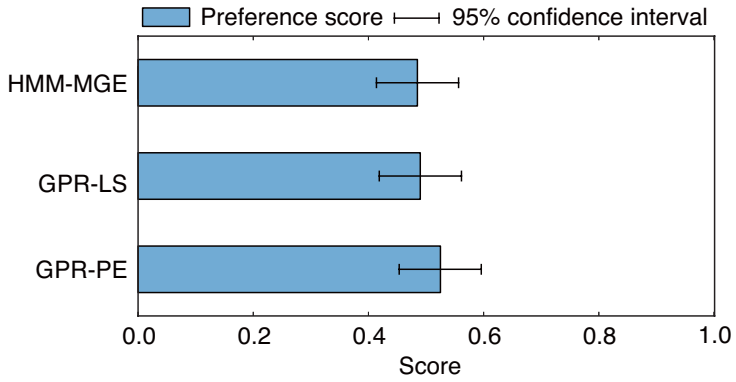


Figure 5.9: Preference score on similarity of synthetic speech to original speech.

PIC and context extension into the proposed technique effectively reduced spectral distortion, and the proposed GPR-based technique significantly improved the naturalness of synthetic speech compared to the conventional HMM-based technique without degrading reproducibility. However, most of the hyper-parameters in this study were manually tuned on the basis of preliminary experimental results. The advantage of GP is that the parameters of kernel functions can be automatically optimized. Therefore, we intend to refine kernel function structures and their parameters in future work to express real data more appropriately and this should leads to improvements in quality.

Chapter 6

Conclusions and Future Work

6.1 Summary of the thesis

This thesis described novel approaches for synthesizing speech with prosodic variability and naturalness.

Chapter 1 described basic background of speech synthesis. There are a variety of applications of speech synthesis and there have been increasing demands for such applications. Although the variability and naturalness of synthetic speech have been improved, the ability of generating natural sounding speech is still insufficient. Then, the scope of thesis and the basic idea of proposed techniques were provided. Chapter 2, described general statistical parametric speech synthesis. This chapter explained HMM-based speech synthesis and how this study explored the framework of statistical parametric speech synthesis.

In chapter 3, extended context was introduced to synthesize natural-sounding spontaneous conversational speech with prosodic variability in HMM-based speech synthesis. Several context sets that can be obtained from the CSJ are introduced and the effectiveness of the context sets are evaluated. The results of objective evaluation show that the phone prolongation and tone labels are effective for improving generated F0 and duration. It was confirmed that the synthetic speech using extended context offers more natural-sounding speech than conventional contexts from the subjective evaluation.

In chapter 4, prosodic-event-based HMM (prosodic-unit HMM) was proposed to improve the naturalness of prosody of spontaneous conversational speech. The modeling unit proposed prosodic-event-based HMM is the segment between two tone labels that represents prosodic events. The results show that the proposed technique gives a more compact model and more variation in generated F0 than phone-unit HMM.

The prosodic variability and naturalness of synthetic speech were improved by extended context and prosodic-event-based HMM. However the naturalness of spectral features is still insufficient. In chapter 5, Therefore, a speech synthesis framework based on Gaussian process regression was proposed to improve the naturalness of spectral features. Block-based sparse GP approximations such as local GPs and PIC were used for trajectory modeling of utterances with feasible computation. Moreover, for the generation of smooth parameter trajectory, frame context including nearby phone information and its kernel is defined. From the objective and subjective evaluation, the proposed method using the PIC approximation and the extended context achieved better performance than the HMM-based methods.

6.2 Future work

Future work will focus on mixing the proposed extended context and prosodic-event-based unit into the speech synthesis of Gaussian process regression. Since Gaussian process regression is a flexible model, it could be expected that the extended context will have more impact on the synthetic conversational speech effectively. In addition, prosodic-event-based unit can be applied to the frame context.

Furthermore, future work should investigate the various kinds of speech like dialog in audiobooks and singing voices. and examine universal modeling techniques that can be applicable for such various kinds of speech.

Bibliography

[1] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalized synthetic voices for individuals with dysarthria using the hts toolkit," *Computer Synthesised Speech Technologies: Tools for Aiding Impairment, IGI Global, Hershey, PA, USA*, pp. 92–115, 2010.

[2] H. Kenmochi, "Singing synthesis as a new musical instrument," in *Proc. ICASSP 2012*, 2012, pp. 5385–5388.

[3] S. King and V. Karaiskos, "The blizzard challenge 2012," *Blizzard Challenge*, 2012.

[4] L. Chen, M. J. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. INTERSPEECH 2012*, 2012.

[5] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] 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*, 1999, pp. 2347–2350.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[8] 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.

[9] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[10] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: an extended J-ToBI for spontaneous speech," in *Proc. 7th ICSLP*, 2002, pp. 1545–1548.

[11] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA, 2006.

[12] A. W. Black, "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling," in *Proc. INTERSPEECH*, 2006.

[13] Z. Heiga, T. Tomoki, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[14] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2006.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. 90, no. 5, pp. 825–834, 2007.

[16] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*. IEEE, 1999, pp. 229–232.

[17] J. J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 1995.

[18] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology*, vol. 21, no. 2, pp. 79–86, 2000.

[19] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, 1995, pp. 660–663.

[20] N. Campbell, "Developments in corpus-based speech synthesis : approaching natural conversational speech," *IEICE Trans. Inf. & Syst.*, vol. 88, no. 3, pp. 376–383, 2005.

[21] T. Akagawa, K. Iwano, and S. Furui, "Toward hidden markov model-based spontaneous speech synthesis," *J. Acoust. Soc. America*, vol. 120, pp. 3037–3038, 2006.

[22] C. Lee, C. Wu, and J. Guo, "Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation," *Proc. ICASSP*, pp. 4826–4829, 2010.

[23] S. Andersson, J. Yamagishi, and R. Clark, "Utilising spontaneous conversational speech in HMM-based speech synthesis," in *Proc. 7th ISCA Workshop on Speech Synthesis*, 2010, pp. 173–178.

[24] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational Spontaneous Speech Synthesis Using Average Voice Model," in *Proc. INTERSPEECH*, 2010, pp. 853–856.

[25] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[26] 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.

[27] Weka 3: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/.

[28] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1-2, pp. 169–186, 1994.

[29] M. Lei, Y. Wu, F. Soong, Z. Ling, and L. Dai, "A hierarchical F0 modeling method for HMM-Based speech synthesis," in *Proc. INTERSPEECH*, 2010, pp. 2170–2173.

[30] Y. Qian, Z. Wu, B. Gao, and F. K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1702–1710, Aug. 2011.

[31] N. Campbell and J. Venditti, "J-ToBI: An intonation labelling system for Japanese," in *Proc. the Autumn meeting of the Acoustical Society of Japan*, vol. 1, 1995, pp. 317–318.

[32] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Second International Conference on Spoken Language Processing*, 1992.

[33] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2263–2266.

[34] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[35] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[36] J. Yu, M. Zhang, J. Tao, and X. Wang, "A novel HMM-based TTS system using both continuous HMMs and discrete HMMs," in *Proc. ICASSP 2007*, 2007, pp. 709–712.

[37] Y. Yan Z.-J., Y. Qian, and F. Soong, "Rich context modeling for high quality HMM-based TTS," in *Proc. INTERSPEECH 2009*, 2009, pp. 1755–1758.

[38] S. Park and S. Choi, "Gaussian process regression for voice activity detection and speech enhancement," in *Proc. IJCNN*, 2008, pp. 2879–2882.

[39] N. Pilkington, H. Zen, and M. Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, 2011, pp. 2761–2764.

[40] G. Henter, M. Frean, and W. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proc. ICASSP*, 2012, pp. 4505–4508.

[41] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[42] T. Fukuda and T. Nitta, "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," *IEICE Trans. Inf. & Syst.*, vol. 87, no. 5, pp. 1110–1118, 2004.

[43] H. Wackernagel, *Multivariate Geostatistics*. Springer, 2003.

[44] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proceedings of Artificial Intelligence and Statistics*, 2007.

[45] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[46] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *NIPS 18, MIT press*, 2006, pp. 1257–1264.

[47] D. Haussler, "Convolution kernels on discrete structures," in *Technical Report UCSC-CRL-99-10*. Dept of Computer Science, University of California at Santa Cruz., 1999.

[48] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 1, 2006, pp. 889–892.

List of Publications

Publications Related to This Thesis

Journal

1. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Statistical nonparametric speech synthesis based on Gaussian process regression,"
   IEEE Journal of Selected Topics in Signal Processing (submitted).

2. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Extension of context set for generating diverse prosodic variations in HMM-based spontaneous conversational speech synthesis"
   IEICE Trans. Information and Systems, vol.J95-D, No.3, pp.597–607 Mar. 2012 (in Japanese).

International Conference

1. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Statistical nonparametric speech synthesis using sparse Gaussian processes,"
   Proc. INTERSPEECH 2013, pp.1072–1076, Aug 2013.

2. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Frame-level acoustic modeling based on Gaussian process regression for statistical nonparametric speech synthesis,"
   Proc. ICASSP 2013, pp.8007–8010, May 2013.

3. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"Discontinuous observation HMM for prosodic-event-based F0 generation,"
Proc. INTERSPEECH 2012, Sept. 2012.

4. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"An F0 modeling technique based on prosodic events for spontaneous speech synthesis,"
Proc. ICASSP 2012, pp.4589–4593, Mar. 2012.

5. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"On the use of extended context for HMM-based spontaneous conversational speech synthesis,"
Proc. INTERSPEECH 2011, pp.2657–2660, Aug. 2012.

IEICE Technical Report

1. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"On the use of prosodic-event-based HMM in F0 generation of conversational speech"
IEICE Technical Report, vol.111, no.365, SP2011-98, pp.185–190, Dec. 2011 (in Japanese).

2. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"Performance evaluation of contexts for conversational speech synthesis using Corpus of Spontaneous Japanese"
IEICE Technical Report, vol.111, no.28, SP2011-27, pp.155–160, May 2011 (in Japanese).

ASJ Meeting

1. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on frame-level acoustic modeling using Gaussian process regression for speech synthesis,"
ASJ Spring meeting, 1-7-5, pp.271–272, Mar. 2013 (in Japanese).

2. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on F0 modeling based on discontinuous observation HMM,"
ASJ Spring meeting, 1-11-6, pp.305–306, Mar. 2012 (in Japanese).

3. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Timing prediction of intonation labels for conversational speech synthesis,"
   ASJ Autumn meeting, 3-8-2 pp.333–334, Sept. 2011 (in Japanese).

4. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "A study on phonetic and prosodic contexts for conversational speech synthesis using Corpus of Spontaneous Japanese,"
   ASJ Spring meeting, 3-7-1, pp.297–298, Mar. 2011 (in Japanese).

Other Publications

International Conference

1. Takashi Nose, Misa Kanemoto, Tomoki Koriyama, Takao Kobayashi,
   "A style control technique for singing voice synthesis based on multiple-regression HSMM,"
   Proc. INTERSPEECH 2013, Aug. 2013.

2. Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,
   "HMM-based expressive speech synthesis based on phrase-level F0 context labeling,"
   Proc. ICASSP 2013, pp.7859–7863, May 2013.

3. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
   "Conversational spontaneous speech synthesis using average voice model,"
   Proc. INTERSPEECH 2010, pp.853–856, Aug 2010.

IEICE Technical Report

1. Takashi Nose, Misa Kanemoto, Tomoki Koriyama, Takao Kobayashi,
   "A study on style control based on multiple-regression HSMM for synthesizing singing voices with various expressivity,"
   IEICE Technical Report, vol.112, no.422, SP2012-111, pp.79–84, Jan. 2013 (in Japanese).

2. Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka,

"An on-line acoustic model adaptation technique based on style estimation," IEICE Technical Report, vol.112, no.422, SP2012-112, pp.85–90, Jan. 2013 (in Japanese).

3. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on conversational speech synthesis based on average voice model," IEICE Technical Report, vol.109, no.375, SP2009-101, pp.33–38, Jan. 2011 (in Japanese).

ASJ Meeting

1. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on conversational speech synthesis based on two-stage model adaptation,"
ASJ Autumn meeting, 2-Q-3, pp.303–304, Sept. 2010 (in Japanese).

2. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on utterance unit for HMM-based conversational speech synthesis,"
ASJ Spring meeting, 3-6-19, pp.143–144, Mar. 2010 (in Japanese).

3. Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
"A study on HMM-based conversational speech synthesis,"
ASJ Autumn meeting, 1-2-10, pp.255–256, Sept. 2009 (in Japanese).