On the Use of Extended Context for HMM-based Spontaneous Conversational Speech Synthesis

Tomoki Koriyama, Takashi Nose, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Japan

koriyama.t.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper addresses an issue of prosodic variability of spontaneous speech in HMM-based spontaneous conversational speech synthesis. We propose an extended context set including information peculiar to spontaneous speech derived from the annotation data embedded in a large-scale database of spontaneous Japanese. We show the effectiveness of the newly introduced contexts from the results of objective and subjective evaluation experiments. We also propose stopping criteria for decision-tree clustering to alleviate an over-fitting problem. Experimental results show that the restriction of the size of each leaf node can improve the quality of synthetic speech.

Index Terms: conversational speech, spontaneous speech, HMM-based speech synthesis, prosodic context, CSJ, X-JToBI

1. Introduction

In practical applications of speech synthesis, the realization of a system that can generate spontaneous speech appearing in human communication is an ultimate goal. 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 sounding speech like a human [1]. Recently, there have been alternative attempts for spontaneous and/or conversational speech synthesis [2–5] using HMM-based synthesis which has shown its advantage in a relatively small amount of training data. In [2], fundamental frequency (F0) contours and phone durations were modeled based on the quantification theory type I. Another prosody modeling technique was proposed in [3], where state-based voice transformation from read speech was used. In [4], a technique based on the multi-space distribution HMM [6], 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 [5].

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. For the expressive speech synthesis, multiple emotions and speaking styles have shown to be well modeled by the style-mixed model with the global context of style types [7]. Emphasis expressions appearing partially in an utterance can be also represented by the context related to the position of the emphasis [8].

In this paper, 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 contexts 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.

2. Extended context for spontaneous conversational speech

In this study, we examine 12 context categories shown in Table 1. These contexts can be determined automatically by the labels annotated in the CSJ core. In the conventional HMM-based speech synthesis, the information 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. 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

Table 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

pitch movements of spontaneous conversational speech are much more complicated than those 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, or 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 inflection points of F0 contour as shown in Fig. 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. Stopping criteria for tree-based context clustering

In the HMM-based speech synthesis, clustering of phonetic and prosodic contexts based on a decision tree is performed in order to improve the reliability of model parameter estimation and

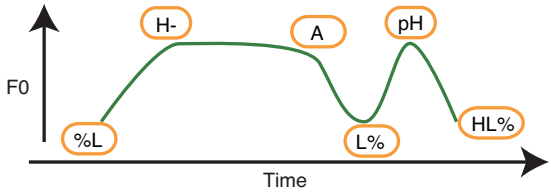

Figure 1: Schematic example of the relationship between X-JToBI tone tier labels and the F0 contour of an accent phrase which ends with rise-fall. Each inflection point is labeled.

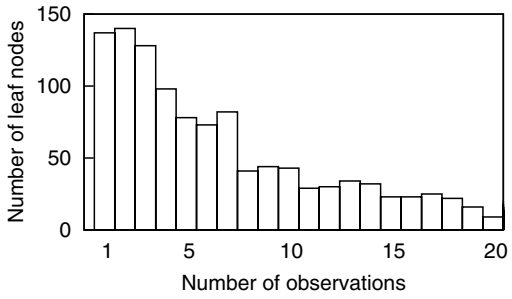

Figure 2: Histogram of the number of observations contained in one leaf node.

to predict parameters for unseen contexts. Each leaf node is split using the question 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 [11]. However, when the extended context described in Sect. 2 is incorporated into the spontaneous conversational speech synthesis, the MDL criterion does not 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 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 here 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.

4. Experiments

4.1. Experimental conditions

We used conversational speech data of two female speakers (ID=19 and 514) included in the CSJ database for evaluation experiments. Each speaker was non-professional speaker and uttered three sets of conversational speech: two interviews and

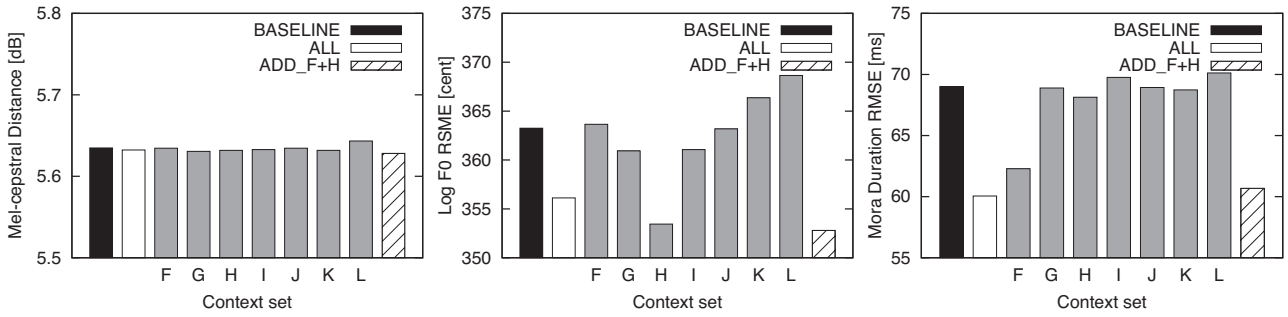

Figure 3: Distortions of acoustic features with different context sets.

a task-oriented dialog. The total length of speech samples of each speaker is approximately 25 minutes. Speech signals were sampled at a rate of 16kHz. The STRAIGHT analysis [12] 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) [13] 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. 10-fold cross-validation tests was performed in the evaluations.

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 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. 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 1 and was added to BASELINE. The results in Fig. 3 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 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.

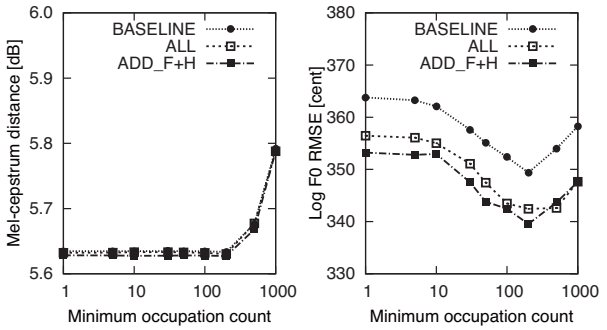

Figure 4: Spectral and F0 distortions as a function of the minimum occupation counts.

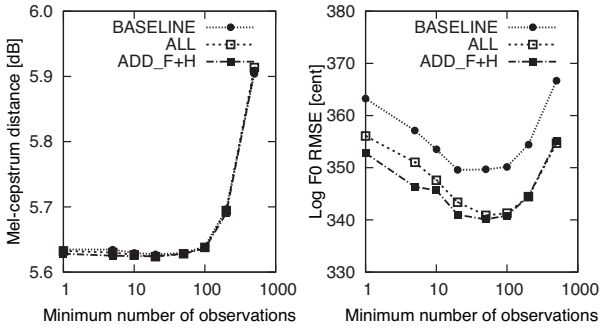

Figure 5: Spectral and F0 distortions as a function of the minimum number of observations.

The results are shown as ADD_F+H in Fig. 3. We can see that the distortions of mel-cepstrum and duration of ADD_F+H were comparable to ALL. Moreover, the F0 distortion of ADD_F+H was slightly lower than that of ALL.

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 4 and 5 show average mel-cepstral distances and RMS errors of log F0. The minimum number of observations is not limited in Fig. 4, and the minimum occupation count is fixed to 5.0 in Fig. 5. As for the mel-cepstral distance, it was not sensitive to the stopping criteria when the minimum number of observations was less than

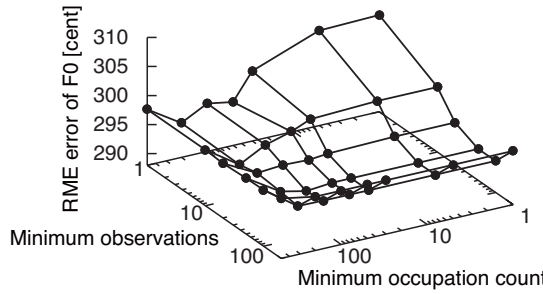

Figure 6: F0 distortions as a function of the minimum occupation count and number of observations.

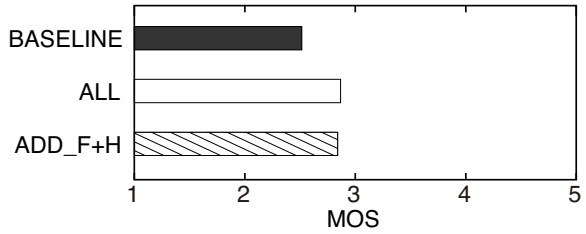

Figure 7: Results of a MOS test on naturalness of synthetic speech.

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 distortions 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 of ADD_F+H was calculated when both of the criteria were used in the clustering. The result is shown in Fig. 6. 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. 4 and 5, the criterion based on the minimum number of observations was less sensitive to distortion variation than that based on the minimum occupation count.

4.4. Naturalness

The naturalness of the synthetic speech was evaluated for the three context sets: BASELINE, ALL, and ADD_F+H by a MOS test. The minimum number of observations was set to 50 in ALL and ADD_F+H based on the above results. Six 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. The average scores are shown in Fig. 7. Both ALL and ADD_F+H gave higher performance than BASELINE, and the difference is statistically significant at the 5% level. ADD_F+H 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.

5. Conclusion

To synthesize spontaneous conversational speech with 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.

6. Acknowledgements

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063 and 23700195.

7. References

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

[2] 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.

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

[4] S. Andersson, J. Yamagishi, and R. Clark, "Utilising spontaneous conversational speech in HMM-based speech synthesis," in *Proc. of 7th ISCA workshop on speech synthesis*, 2010.

[5] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational Spontaneous Speech Synthesis Using Average Voice Model," in *Proc. INTERSPEECH*, 2010.

[6] 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.

[7] 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.

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

[9] Coupus of Spontaneous Japanese http://www.kokken.go.jp/katsudo/corpus.

[10] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: an extended J-ToBI for spontaneous speech," in *Seventh International Conference on Spoken Language Processing*, 2002.

[11] 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.

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

[13] 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.