



Conversational Spontaneous Speech Synthesis Using Average Voice Model

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@ip.titech.ac.jp, takao.kobayashi@ip.titech.ac.jp

Abstract

This paper describes conversational spontaneous speech synthesis based on hidden Markov model (HMM). To reduce the amount of data required for model training, we utilize an average-voice-based speech synthesis framework, which has been shown to be effective for synthesizing speech with arbitrary speaker's voice using a small amount of training data. We examine several kinds of average voice model using reading-style speech and/or conversation-style speech. We also examine an appropriate utterance unit for conversational speech synthesis. Experimental results show that the proposed two-stage model adaptation method improves the quality of synthetic conversational speech.

Index Terms: conversational speech, spontaneous speech, HMM-based speech synthesis, average voice model, speaker adaptation, style adaptation

1. Introduction

Toward practical applications for humanoid spoken dialog systems or aid for speech-impaired people to communicate with speech, there is high expectation for the realization of spontaneous speech synthesis systems [1]. In recent years, although speech synthesis technique enables us to generate smooth and natural sounding speech in reading-style, spontaneous speech synthesis still remains a difficult problem. In fact, so far, there have been reported much fewer studies on spontaneous speech synthesis compared to those on read speech synthesis.

In [2], fundamental frequency (F0) contours and phone durations have been modeled based on quantification theory type I [3], however, it was reported that the generated speech had insufficient spontaneity [4]. In [5], it has been shown that spontaneity was improved by transformation from normal synthetic speech to spontaneous one. This technique needs parallel speech data and this would be impractical for the case of synthesizing arbitrary speakers' conversational speech. Campbell [6] has suggested that the use of a very large conversational speech corpus enables synthetic speech to have a wide variety of expressions. Although this is true, especially for corpus-based speech synthesis including HMM-based system, it is unrealistic to collect a huge amount of spontaneous speech data for respective arbitrary speakers.

In the HMM-based speech synthesis system [7], model training for the speech synthesis units needs not only phonetic and prosodic information but also linguistic one such as syllables, words, phrases, and sentences. A set of such factors, called *context*, plays an important roll in both the training and synthesis stages of the HMM-based speech synthesis. For the spontaneous speech case, automatic phonetic, prosodic and linguistic labeling are much more difficult than the read speech case because of incompleteness in pronunciation and grammar as well as existence of fillers and disfluencies. This leads to difficulty of preparing a sufficient amount of training data of spontaneous speech for arbitrary target speakers. Therefore it is crucial to synthesize speech with acceptable quality using only

a small amount of target speaker's training data for spontaneous conversational speech.

In this paper, we propose an approach to spontaneous conversational speech synthesis based on the HMM-based synthesis framework. We have proposed model adaptation from average voice model [8] as a method of synthesizing arbitrary speaker's speech using a limited amount of the target speaker's training data. Here we utilize the average-voice-based speech synthesis approach and apply it to conversational speech with high spontaneity. We investigate several types of average voice model, specifically, conventional average voice model trained using reading-style speech, average voice models using conversational speech that aims to lessen the difference of characteristics between average voice model and target speaking style. We also examine two-stage model adaptation using both reading- and conversation-style speech. Moreover, we examine the choice of appropriate utterance unit for conversational speech synthesis.

2. Average-voice-based speech synthesis for conversational speech

2.1. Speech synthesis based on average voice model and model adaptation

When the target speaker's speech data is very limited, it is difficult to estimate reliable parameters of target speaker's HMMs. In such a case, model adaptation from an average voice model is an effective way to enable the model training with a small amount of target speaker's data [8]. In this technique, the average voice model is trained in advance using multiple speakers' data, and is adapted to the target speaker's characteristics with a small amount of data. In this study, we apply this technique to conversational spontaneous speech synthesis. For the average voice model, a model trained using reading-style speech has been generally used, which can be applied to the conversation-style spontaneous speech by simultaneous adaptation of speakers and styles [9]. However, conversation- and reading-style speech have considerably different acoustic characteristics, which could be an obstacle to the model adaptation. To alleviate the gap, the average voice model trained using conversational spontaneous speech would be more preferable.

2.2. Two-stage model adaptation for conversational speech synthesis

Although the model adaptation from the conversation-style average voice model is more suitable in terms of the acoustic similarity, it generally takes high cost to prepare a sufficient amount of training data of multiple speakers' spontaneous speech as well as the target speaker's data. To relax the data sparseness and acoustic gap, here we propose a two-stage model adaptation technique as shown in Fig. 1. In this technique, we first train an average voice model using a sufficient amount of reading-style speech data, and conduct the style adaptation to the conversation-style using a small amount of conversational

10.21437/Interspeech.2010-190

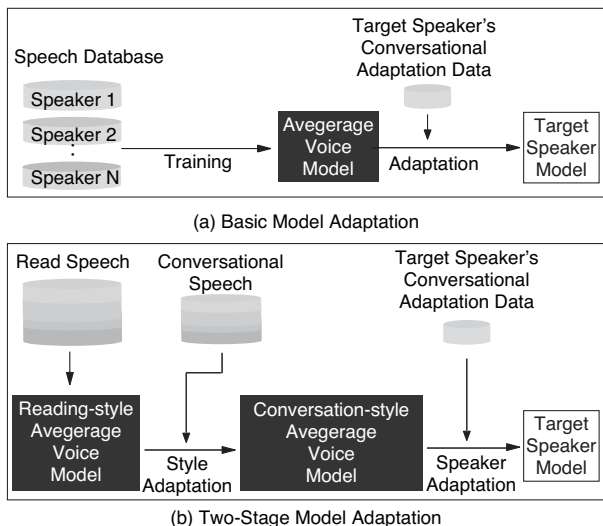


Figure 1: Basic and two-stage model adaptation procedures from average voice model.

speech data of multiple speakers. Then, we obtain the target speaker’s conversation-style model by adapting the speaker individuality of the conversation-style average voice model to that of the target speaker. By means of the two-stage adaptation, we expect to lessen the difference between average voice model and target speech, and to reduce the cost for collecting conversational spontaneous speech.

3. Utterance unit for conversational speech

In the HMM-based speech synthesis, the determination of utterance units is one of the important issues. This is because a number of prosodic contextual factors for each phone, e.g., position of the current phrase in an utterance, are used for the modeling of context-dependent synthesis units, and these factors depend on the utterance unit. In conventional reading-style speech synthesis systems, the utterance unit is explicitly determined by a formal sentence given in recoding. By contrast, the conversational spontaneous speech is generally recorded session by session with a certain topic. It is sometimes difficult to determine the sentence because the expression of the sentence-final is often omitted, and there are a lot of changes of the conversational turn. To determine an appropriate utterance unit for the alternative to a sentence, we examine what kind of unit is perceived as natural.

3.1. Conversational spontaneous speech database

For the conversation-style speech, we used a Japanese spontaneous speech database, *the corpus of spontaneous Japanese* (CSJ) [10]. We chose three females (ID:19, 463, and 514) and three males (ID: 423, 471, and 685). They were not professional narrators, and each of them uttered three sets of conversational speech: two interviews and a task oriented dialog. Each conversational speech was transcribed and annotated, and there are a lot of annotations such as part of speech, clause unit, and intonation information. Here the clause is a grammatical unit that consists of a subject and a predicate, and the clause boundaries were automatically determined by transcription data.

3.2. Determination of utterance unit based on silences

One of the cues for the determination of utterance unit is a silence in the conversation. Here, we focus on the distribution of the length of silences in the conversational speech. We com-

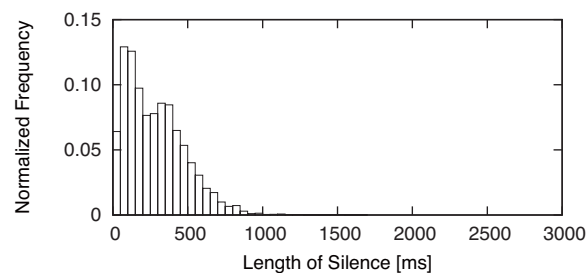


Figure 2: Histogram of pause length in reading-style speech.

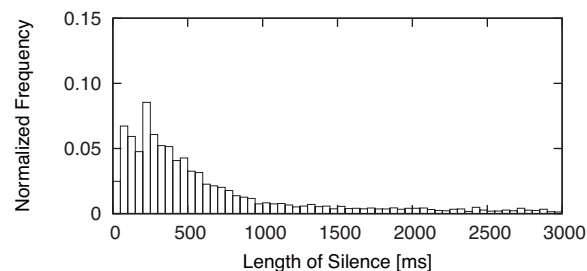


Figure 3: Histogram of silence length including pause in conversation-style speech.

pared the histograms of reading- and conversation-style speech. Figures 2 and 3 show the results. For the reading-style speech, ATR Japanese database set B [11] was used where ten professional speakers uttered 503 phonetically balanced sentences, 5030 sentences in total. In the reading-style speech, there are almost no silences over 1500ms. However, there are longer silences than 1500ms in conversation-style speech. This indicates that there are two types of silence, not only the silence as a pause in utterances, but also the silence as a period when the speaker is listening to another speaker by turn-taking or thinking about what to say next. From Fig. 3, we defined that 1500ms or longer silences are those that are not pauses in an utterance. We used this definition in the following experiment.

3.3. Choice of utterance unit for conversational speech

By taking account of the above discussion, we consider three kinds of utterance unit for conversational speech. The first utterance unit is a portion segmented by 200ms or longer silences. Although this unit is used as the basic unit of transcription in CSJ, the resultant segmented utterances are relatively short, and there are a lot of fragmented speech including only one word. The second one is a portion segmented by 1500ms or longer silences that can be regarded as a boundary of pause and non-pause silence from the previous experimental results. The third one is a combination of the second one and the clause, that is, segmentation by 1500ms or longer silences and clause boundaries. When using the second utterance unit, some utterances became too long and were perceived as unnatural. We expect that the use of the clause information can segment such utterances appropriately and make utterance units more natural.

3.4. Perceptual naturalness of segmented speech

We evaluated naturalness of speech samples of CSJ segmented by the three kinds of utterance unit described in Sect. 3.3. Five participants listened to the speech samples, and judged whether the utterance sounded “natural” or “unnatural.” Each participant evaluated randomly chosen 60 samples for each utterance unit. The result of all speakers’ total is shown in Fig. 4. The score is the rate of the utterances evaluated as “natural” or “unnatural” to

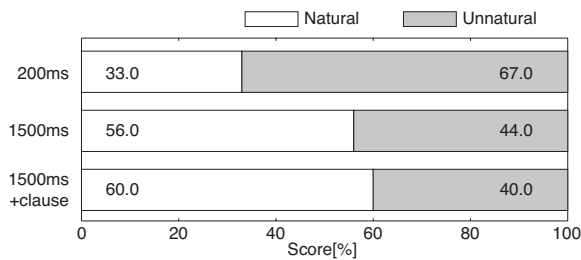


Figure 4: Preference scores of naturalness of segmented conversational speech with different utterance units.

the total of evaluated utterances. From the result, it can be seen that the utterances segmented with 200ms or longer silences have low naturalness, and that the utterances segmented with 1500ms or longer silences and clause boundary have higher naturalness than the others. Therefore we use the utterance unit determined by the clause boundaries with 1500ms or longer silences in the following experiments.

4. Experiments

4.1. Experimental conditions

Speech signals were sampled at a rate of 16kHz and windowed with a 5ms shift. The feature vector consisted of 25 mel-cepstral coefficients including zeroth coefficient obtained by STRAIGHT [12] 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 HSMMs without skip paths. Each state had a single Gaussian distribution with a diagonal covariance matrix. For training and testing, we used the phonetic and prosodic context labels automatically converted from the labels given in CSJ database.

For the training of the reading-style average voice model, we used professional narrators of ATR Japanese database set B. We examined the following four types of average voice model:

CONV50 Average voice model trained with 5 speakers' conversational speech of 10 minutes for each speaker, 50 minutes in total. The five speakers used in the training were different from the target speakers.

READ50 Average voice model trained with 5 speakers' reading-style speech of 10 minutes for each speaker, 50 minutes in total, that is, almost the same amount as CONV50. The speakers were 3 males (MHO, MMY, and, MSH) and 2 females (FTK and FYM).

READ240 Average voice model trained with 8 speakers' reading-style speech of 30 minutes for each speaker, 240 minutes in total, where we assumed that there was a lot of reading-style speech data available for the model training. The speakers were 4 males (MHO, MMY, MSH, and MYI) and 4 females (FKN, FKS, FTK, and FYM).

TWO-STAGE Conversation-style average voice model obtained by two-stage model adaptation. For reading-style speech, we used the same one as in READ240, and for conversational speech data of multiple speakers, we used the same one as in CONV50.

In the training stage of the average voice models, the shared-decision-tree-based context clustering (STC) algorithm [14] and the speaker adaptive training (SAT) [15] were applied to normalize the influence of speaker differences among the training speakers. In the speaker adaptation, we used the combination of constrained structural maximum a posteriori linear regression (CSMAPLR) and MAP adaptation [8]. In the performance evaluation, we used one female (ID: 463) and one male (ID: 471)

speakers in CSJ. We performed 2-fold cross-validation tests using utterances of 5 minutes that were not included in the adaptation nor training data¹.

4.2. Objective evaluation

To objectively assess the proposed technique, we calculated the distortions of generated spectrum, F0, and phoneme duration of synthetic speech against those of target speaker's original speech. Figures 5 shows the average cepstral distance, root mean square (RMS) error of log F0, phoneme duration, respectively, when changing the amount of training data from 1 minute to 5 minutes.

By comparing READ50 and CONV50 which had almost the same amount of training data for average voice model, we can see that CONV50 gave much better performance than READ50. This implies that the acoustic similarity of the speaking styles and spontaneity improves the adaptation performance. However, the performance of the reading-style average voice model was also improved by increasing the amount of training data, and a rich amount of contexts in the training data is effective in the reproducibility of acoustic features even if the styles of training data is different. Moreover, it was found that the adaptation performance of READ240 was enhanced by using the two-stage model adaptation.

4.3. Subjective evaluation

We evaluated subjectively the naturalness and spontaneity of the synthetic speech of the three methods, READ240, CONV50, and TWO-STAGE. We generated speech samples from the adapted models using 5 minutes utterances. Six participants were requested to make a rating from five choices: 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. Each participant evaluated 20 utterances for each method, randomly chosen from synthetic speech samples used for the objective evaluation.

The average scores of all votes in mean opinion score (MOS) are shown in Fig. 6. The error bar represents a 95% confidence interval of each score. The objective evaluation results showed that CONV50 gave the best performance in the reproducibility of the spectral and duration features, whereas the naturalness and spontaneity of CONV50 were the lowest of all three methods. A possible reason is that the F0 distortion in the case of CONV50 was worst of three methods, and this caused crucial degradation of the perceptual naturalness. In contrast, the performance of READ240 was better than CONV50 because of the sufficient amount of training data. Moreover, although there is not a significant difference of naturalness between READ240 and TWO-STAGE, the average score was slightly improved by using the two-stage model adaptation.

5. Conclusion

This paper presented the choice of an utterance unit for conversational speech and a technique synthesizing conversational spontaneous speech under condition where target speaker's available speech is limited. From the experimental results, the utterance units segmented by 1500ms or longer silences and the clause boundaries were perceived more natural as a conversational utterance unit. A two-stage model adaptation technique, style adaptation from reading-style to conversation-style and speaker adaptation from average voice to target speaker was proposed and evaluated objectively and subjectively. From the evaluation results, the two-stage model adaptation method can improve the naturalness to the method using the average voice models trained by only reading-style speech and only conversational speech, respectively.

¹Several speech samples used in the test are available at <http://www.kbys.ip.titech.ac.jp/demo/is2010/koriyama/>

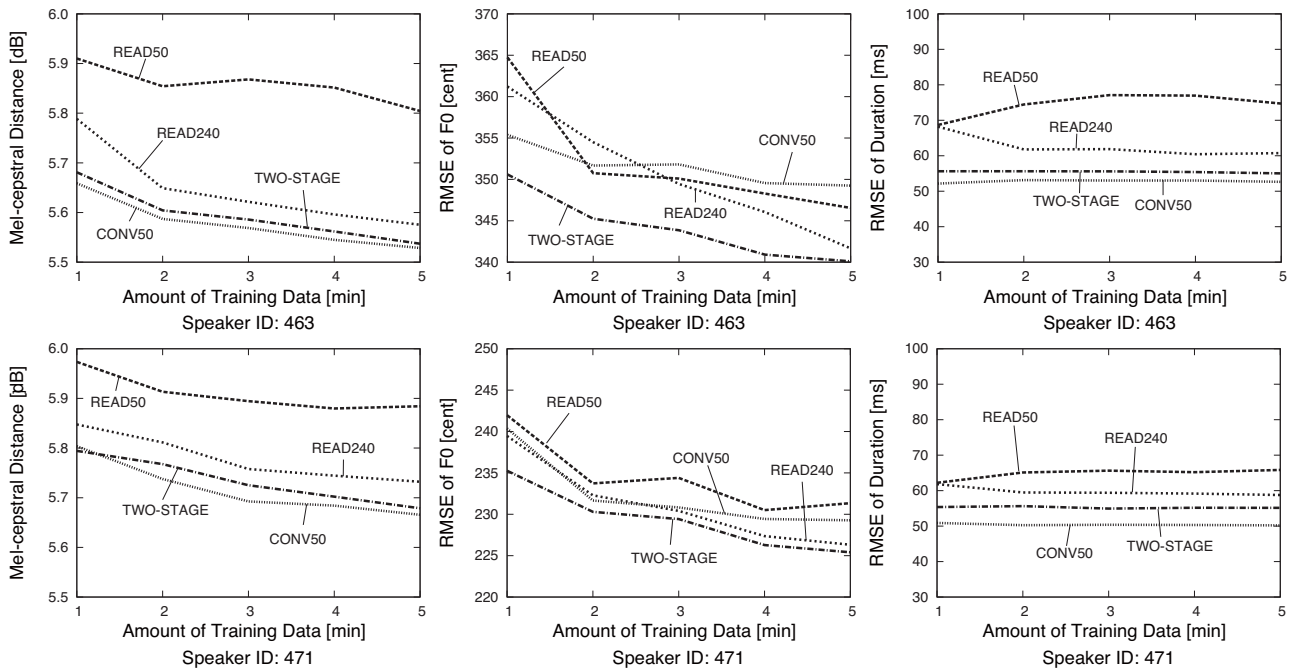


Figure 5: Distortion of spectrum, F0, and phoneme duration between synthetic and original speech using different types of average voice model.

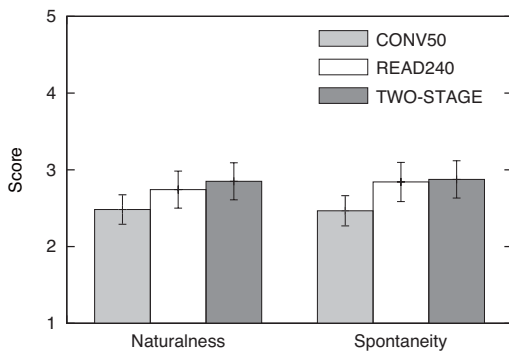


Figure 6: MOS Listening scores for synthetic conversational speech.

6. Acknowledgements

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063 and 21800020.

7. References

- [1] S. Werner, M. Eichner, M. Wolff, and R. Hoffmann, "Toward spontaneous speech synthesis-utilizing language model information in TTS," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 4, pp. 436–445, 2004.
- [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] K. Iwano, M. Yamada, T. Togawa, and S. Furui, "Prosody control for HMM-based Japanese TTS," *Text to speech synthesis: new paradigms and advances*, p. 155, 2005.
- [4] T. Akagawa, K. Iwano, and S. Furui, "A study on the Statistical models for HMM-based spontaneous speech synthesis," *IEICE technical report (in Japanese)*, vol. 107, no. 77, pp. 13–18, 2007.
- [5] C. Lee, C. Wu, and J. Guo, "Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation," *INTERSPEECH*, 2010.
- [6] N. Campbell, "Developments in corpus-based speech synthesis : approaching natural conversational speech," *IEICE Trans. Inf. & Syst.*, vol. 88, no. 3, pp. 376–383, 2005.
- [7] 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*, Sept. 1999, pp. 2347–2350.
- [8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Iso-gai, "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.
- [9] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, "Speaker and style adaptation using average voice model for style control in hmm-based speech synthesis," in *ICASSP*, 2008.
- [10] Coupus of Spontaneous Japanese <http://www.kokken.go.jp/katsudo/corpus>.
- [11] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, Aug. 1990.
- [12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Re-structuring 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, Apr. 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.
- [14] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol. 86, no. 3, pp. 534–542, 2003.
- [15] T. Anastasakos, "A compact model for speaker-adaptive training," *ICSLP*, vol. 2, 1996.