Fluent Personalized Speech Synthesis with Prosodic Word-Level Spontaneous Speech generation

Yi-Chin Huang, Chung-Hsien Wu, Ming-Ge Shie

Dept. of Computer Science and Information Engineering
National Cheng-Kung University, Tainan, Taiwan
ychin.huang@gmail.com, chunghsienwu@gmail.com

Abstract
This paper proposes an automatic approach to generating speech with fluency at the prosodic word level based on a small-sized speech database of the target speaker, consisting of read and fluent speech. First, an auto-segmentation algorithm is employed to automatically segment and label the database of the target speaker. A pre-trained average voice model is adapted to the voice model of the target speaker by using the auto-segmented data. For synthesizing fluent speech, a prosodic model is proposed to smooth the prosodic word-level parameters to improve the fluency in a prosodic word. Finally, a postfilter method based on the modulation spectrum is adopted to alleviate over-smoothing problem of the synthesized speech and thus improve the speaker similarity. Experimental results showed that the proposed method can effectively improve the speech fluency and speaker likeness of the synthesized speech for a target speaker compared to the MLLR-based model adaptation method.

Index Terms: prosodic word-level smoothing, fluent speech synthesis, personalized speech synthesis

1. Introduction

Spontaneous speech, different from read speech, is highly speaker-dependent, which includes more dynamic speaking rate and less articulated phonemes between nearby syllables in Mandarin. Several research studies focused on synthesizing spontaneous speech [1] [2] [3] lately. The main differences between spontaneous speech and read speech are the speaking rate and the construction of prosodic units. As mentioned in the literature [4] [5], the number of syllables of read speech is smaller than that of spontaneous speech at the same prosodic unit level.

Several researchers focused on modeling pronunciation variation for synthesizing the spontaneous speech or expressive speech [6]. A prediction method is proposed to obtain pronunciation category and build a classification and regression tree (CART) for selecting a suitable pronunciation sequence of the input text [7].

Another method is to construct a pronunciation network according to the result from a speech recognizer for pronunciation sequence selection [8]. These methods require experts to construct a delicate pronunciation dictionary or language model manually since the phoneme sequences of pronunciation variation do not generally exist in the read speech corpus. These studies usually require manual labeling work, which is time consuming and could be error-prone. Several methods, such as HMM [9], CRF [10], BIC [11] have been adopted for automatic segmentation. HMM is trained by the labeled corpus to segment the speech based on forced alignment. Model adaptation based on HMM was used to match the target speech corpus to improve the segmentation accuracy. SVM is trained for refining the boundary of phonemes [9]. Another way is to define a cost function using various acoustic features for speech segmentation [12].

In this study, to achieve the goals for synthesizing a personalized fluent speech with practical situation, only a small-sized database without manual label is required. First, we perform automatic segmentation on a read speech corpus from the target speaker to find the segmented phone sequence. Model adaptation is then adopted to adapt an average voice model (AVM) [13] to the personalized voice model of the target speaker using the small-sized corpus of read speech of the target speaker. For the following fluent speech generation, the Mel Generalized Cepstral (MGC) coefficient sequence of adjacent syllables within a basic prosodic unit are overlapped and smoothed to generate a smooth transition. The best overlapping ratio is decided automatically based on the minimum distance between the overlapped segment and its corresponding spontaneous speech segment. The overlapping ratio will be stored in a decision tree, which is then used in the synthesis phase for synthesized parameter smoothing. Fig. 1 shows the training procedure of the proposed method.

2. Personalized Fluent Speech Generation

For personalized speech synthesis, the source model, constructed from the read speech corpora of several speakers, is
adapted using an automatically segmented small-sized read speech corpus of the target speaker to synthesize the personalized speech. The main aspect of the personalized speech synthesis is that the design of the collected corpora has to follow the criterion: the read speech corpus of the target speaker should consist of all the basic phonemes of Mandarin to maintain the intelligibility of the synthesized speech. On the other hand, the spontaneous speech corpus of the target speaker should consist of as many combinations of two phonemes as possible for better modeling the fluent transition between two phonemes.

2.1. Model Definition and Automatic Phone Segmentation

In the constructed Mandarin speech synthesis system, each Mandarin syllable is composed of at most three phones which are used as the basic synthesis units.

\[ C + V1 + V2, \]

(1)

The first phone, \(C\), is an extended initial portion of a syllable, and is followed by two final tonal phones, \(V1\) and \(V2\). The initial phone \(C\) is generally a consonant, and the phones \(V1\) and \(V2\) compose the rhyme part of a syllable. The five lexical tones of a Mandarin syllable are encoded as the combination of high/middle/low (H/M/L) categories of phones in the rhyme part, according to the lexical tone of the final portion. The lexical tones can be represented as Tone 1 (high): HH, Tone 2 (rising): LH, Tone 3 (low): LL, Tone 4 (falling): HL, and the neutral tone: MM [14]. The training and synthesis procedures follow the framework proposed in [15] [16].

For the unsegmented corpora, the forced alignment method is used for initial phone boundary detection. Specifically, the general phone HMMs are trained using a Mandarin read speech database (TCC300 [17]) using HTK toolkit [18]. To achieve a better segmentation result of the spontaneous speech, a robust database (TCC300 [17]) using HTK toolkit [18]. To achieve a better segmentation result of the spontaneous speech, a robust database (TCC300 [17]) using HTK toolkit [18].

2.2. Prosodic Word-Level Smoothing

In spontaneous speech, the speech flow is more fluent and continuous than that of read speech, and thus results in indistinguishable boundaries between syllables, especially within a prosodic word. Furthermore, spontaneous speech may consist of pronunciation variation because of its dynamic speaking rate. We propose a method to address these issues by smoothing the acoustic parameters between phones within a prosodic word and thus improve the continuity to make the synthesized speech sounds more fluent. In this study, prosodic word prediction is achieved by a CRF model [24]. As Fig. 2 shows, after phone segmentation, the Dynamic Time Warping (DTW) algorithm is used to decide the overlap region which achieve the minimal Euclidean distance of the cepstral coefficients between spontaneous speech and synthesized speech. Then the smoothing ratio is calculated according to the overlap region. The ratio of the length of spontaneous part and the corresponding synthesized part after parameter smoothing is also stored. These ratios are calculated as follows:

\[ i = \arg \min_i C(i) \]

(5)

\[ C(i) = 1/Q \| f_s - fr \| \]

(6)

\[ \text{SmoothRatio} = i/T, \]

\[ \text{LengthRatio} = L_i/L_s \]

(7)

In DTW procedure, \( C(i) \) is the cost function to decide the length \( i \) of the overlapped region. It calculates the Euclidean distance between the overlapped frames and the aligned frames of spontaneous speech, \( fr \) and \( f_s \) respectively. \( Q \) is the cepstral feature dimension. After DTW procedure, the optimal value \( i \) can be obtained and SmoothRatio can be calculated using Eq. (7), where \( T \) is the length of the overlapped region.
In order to mimic the speaking rate of the spontaneous speech, LengthRatio is the ratio of the length of the overlapped synthesized speech $L_s$ to that of the corresponding spontaneous speech, $L_s$.

### 2.3. Decision Tree-based Clustering for Parameter Smoothing Ratio

To generalize the smoothing ratio for similar contextual information of a phone sequence, we adopt the tree-based clustering method to cluster the smoothing ratios of similar contextual phonetic information. The phonetic information adopted as the clustering cues includes the phone identity and articulatory features of the phone. We use tri-gram information consisting of last vowel ($SubL$) of the preceding syllable, and the first consonant ($SubR$) and first vowel ($SubRR$) of the current syllable, respectively, as shown in Fig. 3. Initial clustering is performed according to the manner of articulation of the $SubL$ and $SubR$ because their interactions will affect the pronunciation variation. Then, further clustering based on tri-gram information, including their phone identities or place of articulation, is performed. The clustering criterion is based on minimum description length (MDL) [23] to select the optimal question that splits the cluster with the minimum description length.

### 2.4. Synthesis Stage with Postfiltering

Traditional voice generation using model-based method usually generates over-smoothed parameter trajectories [25], as the upper part of Fig. 4 shows. The over-smoothed parameters of the synthesized speech degrades the naturalness of speech, and therefore the generated speech could be dissimilar to the target speaker. Therefore, postfiltering-based method has been proposed to solve this problem [26]. Here, we adopt the idea and try to apply it to the generated features of the synthesized spontaneous speech in our system. Modulation spectrum (MS), which is defined as a value calculated using the second discrete Fourier transform (DFT) coefficient of a given spectral band, across frame index, with discrete frequency fixed, to capture the spectral changes in that band with different rates [27], is used to calculate the difference between cepstral coefficients of spontaneous speech and read speech. As shown in the lower part of Fig. 4, there are some differences between read speech and spontaneous one, especially in the high frequency range, which is over-smoothed in the synthesized speech. We implemented the postfiltering procedures based on the work proposed by [27]. In the training stage, the MS pairs are constructed by the pseudo-parallel sentences of spontaneous speech of the target speaker and the corresponding synthesized speech, which are adopted for training the GMM-based transformation functions. The decision tree-based clustering for GMM-based transformation functions are then constructed for transforming the out-of-corpus synthesized speech. Here, we use the number of syllables of the utterance as the questions for tree splitting. In the synthesis stage, in order to transform the synthesized speech from the over-smoothed speech to a natural one, first MS is calculated from the synthesized speech and then the most similar GMM pair from the decision tree is retrieved to transform the synthesized speech for compensating the differences. The original phase of the synthesized speech is used and then the inverse Fourier transform is performed to obtain the cepstral coefficient of the target speaker.

### 3. Evaluation

#### 3.1. Speech Data Collection and System Setup

For evaluating the proposed system, several corpora were used. In the proposed method, an A VM is required as the source model. We used TsingHua-Corpus of Speech Synthesis (TH-CoSS) [28] as the main corpus for constructing the A VM. There are two speakers, including one male and one female speaker, in the corpus and 2000 labeled utterances for each speaker were used. Besides, we also collected speech data of four native speakers (two males and two females) with the same utterances as the ones in TH-CoSS. For constructing the target speaker voice model, two speech corpora are required, including read speech and spontaneous speech of the target speaker. Here, both corpora were not segmented. The read speech corpus was designed for covering the phonetic coverage, which consists of all Mandarin phonemes. The spontaneous speech corpus was designed for including various phoneme combinations to preserve the pronunciation variations (PVs). The recorded speech in the target speaker voice model contains conversations in the conference and some acted spontaneous speech with PVs that did not utter in the collected conversations. To make voice less acted, we asked the subjects to record 3 versions for each sentence and chose the most natural one. The statistics of the corpus shows the speaking rate of the spontaneous speech is faster than the read speech, and its variation is also more dynamic.

The MLLR-based adaptation system is served as the “baseline” system. The baseline system was also adapted from the A VM using the spontaneous speech automatically segmented by the proposed method. For the proposed method, we adapted the A VM using the read speech data also automatically segmented by the proposed method, and then applied the proposed PW-level parameter smoothing. The reason for choosing read speech data as the adaptation data is that read speech is more clearly articulated by the speakers, which helps construct a bet-
ter quality voice model, while speech fluency is inferior than the baseline system. This system is denoted as “smooth.” The second proposed system is the proposed system “smooth” with postfiltering, denoted as “smooth_filter.” Besides the proposed systems, a system which used the same adapted model as system “smooth” while only increased or decreased the speaking rate based on the statistics in the collected corpus. This system was used to see if speaking rate change is good enough for generating spontaneous speech, and this system is denoted as “speed.”

3.2. Segmentation Accuracy

For evaluating the proposed method, we firstly compared the segmentation results with the proposed automatic segmentation and the forced alignment using HMM. In the experiment, the manually labeled 200 sentences in the TH-CoSS were used. Given the manually-labeled frame sequence and the automatically-labeled frame sequence, the frame-level segmentation accuracy was used as the evaluation metric. The results are shown in Fig. 5. The proposed automatic segmentation method achieved more precise segmentation results than the forced alignment, which proves that using AF to refine the segmentation result from forced alignment improve the segmentation accuracy effectively.

3.3. Listening Tests for Speech Fluency and Speaker Similarity

The subjective listening tests were conducted to evaluate whether the proposed parameter smoothing with postfiltering improves speech fluency and speaker similarity. Ten native Mandarin subjects participated in the listening tests and 20 outside testing sentences were used for evaluation. The testing sentences were generated as the same collected spontaneous speech uttered by the target speaker with exactly the same length of each syllable.

The 5-point mean opinion score (MOS) test was conducted to evaluate the speech spontaneity. The result is shown in Fig. 6. Surprisingly, the system “speed” is the worst one. With only changing the speaking rate, it can not capture the smooth transition between phones, instead it generates awkward synthesized speech with rapidly uttered syllables or lengthened articulated phones. The “baseline” system generates fluent speech generally. However, there are some inevitably noise caused by poorly adapted models using adaptation data with short duration. The proposed systems generate speech with smooth transition. There is no significant differences between the MOS results of two proposed systems, but there are some feedbacks regarding better speech quality and naturalness for the system “smooth_filter”.

We further evaluated the speaker similarity of the three systems without including system “speed”. The ABX tests were conducted to compare each pair of the comparing systems. The same set of the testing sentences was used. Some reference waveforms of the target speaker were presented to the subjects for judging which synthesized speech in the two comparing systems is more like the sentences uttered by the target speaker. Fig. 7 shows the ABX results. Among the three systems, the synthesized speech from “smooth_filter” was judged the optimal one (significantly differences among three systems). The “smooth” is better than “baseline” with statistical significance. Thus postfiltering is helpful for compensating the over-smoothing problem and generating more fluent synthesized speech.

4. Conclusions

In this paper, we proposed a method for automatic segmentation and a method for prosodic word-level smoothing to realize a fluent personalized TTS system with less manual intervention, with only a small-sized corpus of the target speaker. From the experimental results, prosodic word-level smoothing is effective to improve the speech fluency in a prosodic word; AF is effective to refine automatic segmentation results; and the modulation spectrum-based postfilter improves speaker similarity of the generated speech.
5. References


