PAPER

Improving Naturalness of HMM-Based TTS Trained with Limited Data by Temporal Decomposition

Trung-Nghia PHUNG[†a)], **Thanh-Son PHAN**[††b)], **Thang Tat VU**[††c)], **Mai Chi LUONG**[††d)], ***Nonmembers***, ***and*** **Masato AKAGI**[†e)], ***Member***

SUMMARY The most important advantage of HMM-based TTS is its highly intelligible. However, speech synthesized by HMM-based TTS is muffled and far from natural, especially under limited data conditions, which is mainly caused by its over-smoothness. Therefore, the motivation for this paper is to improve the naturalness of HMM-based TTS trained under limited data conditions while preserving its intelligibility. To achieve this motivation, a hybrid TTS between HMM-based TTS and the modified restricted Temporal Decomposition (MRTD), named HTD in this paper, was proposed. Here, TD is an interpolation model of decomposing a spectral or prosodic sequence of speech into sparse event targets and dynamic event functions, and MRTD is one simplified version of TD. With a determination of event functions close to the concept of co-articulation in speech, MRTD can synthesize smooth speech and the smoothness in synthesized speech can be adjusted by manipulating event targets of MRTD. Previous studies have also found that event functions of MRTD can represent linguistic information of speech, which is important to perceive speech intelligibility, while sparse event targets can convey the non-linguistics information, which is important to perceive the naturalness of speech. Therefore, prosodic trajectories and MRTD event functions of the spectral trajectory generated by HMM-based TTS were kept unchanged to preserve the high and stable intelligibility of HMM-based TTS. Whereas MRTD event targets of the spectral trajectory generated by HMM-based TTS were rendered with an original speech database to enhance the naturalness of synthesized speech. Experimental results with small Vietnamese datasets revealed that the proposed HTD was equivalent to HMM-based TTS in terms of intelligibility but was superior to it in terms of naturalness. Further discussions show that HTD had a small footprint. Therefore, the proposed HTD showed its strong efficiency under limited data conditions.

key words: *text to speech, HMM-based TTS, hybrid TTS, limited data, temporal decomposition*

1. Introduction

Building a huge speech corpus is a costly task that takes a long time and requires a great deal of effort by engineers, acousticians and linguists. Therefore, how to build high-quality TTS under limited data conditions is important for practical speech applications, especially for under-resourced languages.

Two current state-of-the-art TTSs are unit selection and HMM-based TTSs. Unit selection is the most natural-sounding TTS at present. However, since unit selection requires huge data for concatenation, it is a difficult challenge to use it under limited data conditions. HMM-based TTS has been widely studied for the two last decades [1]–[6]. The spectral and prosodic features of speech are modeled and generated in this approach in a unified statistical framework using HMMs. A decision-tree based context clustering technique has been used to ensure the synthesized trajectory is smooth and stable with a limited amount of training data [1]. Therefore, the intelligibility of synthesized speech is still high even under limited data conditions. HMM-based TTS can simultaneously transform the voice characteristics of synthetic speech into those of a target speaker using a small amount of target data by utilizing "average-voice-based" methods [2]. Therefore, it is flexible to adapt the synthetic voice into different target individual voices or target speaking styles with limited amounts of target data. As its trained statistical parameters are small, HMM-based TTS has a small footprint. The runtime computational load of HMM-based TTS is also low. As a result, HMM-based TTS can be easily distributed on different hardware platforms. Although HMM-based TTS has many advantages as was previously mentioned, HMM-based TTS is still far from natural, which is mainly due to buzziness and over-smoothness in synthesized speech. The former is a common issue with speech coding, which has recently been significantly improved, while the latter is caused by "averagely" statistical processing in HMM-based TTS. Although both the spectral and prosodic trajectories generated by HMM-based TTS are over-smooth, the effect of over-smoothness in a spectral sequence is more serious due to the complexity of spectral features.

Many studies have attempted to solve over-smoothness in HMM-based TTS. Using multiple mixtures for modeling state output probability density can reduce over-smoothness in synthesized speech [3]. However, these methods cause another problem with over-training due to the increased number of model parameters. A method of combining continuous HMMs with discrete HMMs, and a method of increasing the number of HMM states has also reduced the over-smoothness in HMM-based TTS [4]. However, these methods increase the complexity of HMMs and are not convenient in practical synthesis systems. The state-of-the-art method to reduce over-smoothness is the parameter generation algorithm that take into consideration global variance

Manuscript received April 8, 2013.
Manuscript revised July 15, 2013.

[†]The authors are with the Japan Advanced Institute of Science and Technology (JAIST), Nomi-shi, 923–1292 Japan.
[††]The authors are with the Institute of Information Technology, IoIT, Vietnam.
a) E-mail: ptnghia@jaist.ac.jp
b) E-mail: ptson@ioit.ac.vn
c) E-mail: vtthang@ioit.ac.vn
d) E-mail: lcmai@ioit.ac.vn
e) E-mail: akagi@jaist.ac.jp

DOI: 10.1587/transinf.E96.D.2417

(GV) [5]. Parameters are generated in this method based on criteria of not only maximizing the HMM likelihood for static and dynamic features but also the likelihood for GV. The experimental results with this method revealed that the naturalness of synthetic speech was significantly improved [5]. However, over-smoothness was still considerable.

Over-smoothness in HMM-based TTS is mainly affected by the accuracy of model estimates [4]. This factor is affected by the amount of training data [6]. The larger the amount of training data, the more accurate the model estimates, and the lesser the over-smoothness in synthesized speech. As a result, the effect of over-smoothness becomes more serious in a situation with limited training data. Therefore, it is difficult to ensure the naturalness of HMM-based TTS under limited data conditions.

Hybrid approaches between HMM-based TTS and unit selection, such as HMM trajectory tiling TTS (HTT) [7], have recently been studied as another solution to improve the naturalness of HMM-based TTS and to preserve the high intelligibility of HMM-based TTS. The HMM trajectory is used to guide the selection of each 5-ms frame to concatenate the waveforms in HTT. The naturalness of HTT is comparable to that of unit selection TTS and its intelligibility is comparable to that of HMM-based TTS. Additionally, this TTS is language-independent due to the use of short frames instead of phonetic-level units. However, HTT still has drawbacks. The major one is the use of short frames, which requires a perfect selection process. If the selection process is imperfect due to a limited data corpus, it may be easy to perceive discontinuities between frames. As a result, this TTS still requires a huge amount of data for rendering. Additionally, HTT is not able to preserve other advantages of HMM-based TTS such as its flexibility for voice adaption and its small footprint.

Based on the above considerations, the motivation for this paper is to improve the naturalness of HMM-based TTS under limited data conditions while preserving its intelligibility. To achieve this motivation, a hybrid TTS between HMM-based TTS and MRTD [9], named HTD in this paper, was proposed. Here, TD is a sparse interpolation model that decomposes a spectral or prosodic sequence into two mutually independent components: static event targets and corresponding dynamic event functions [8], and MRTD [9] is one compact but efficient version of TD with a small interpolation error. With a determination of smooth event functions close to the concept of co-articulation in speech, MRTD can synthesize smooth speech and the smoothness in synthesized speech can be adjusted by modifying event targets of MRTD. Therefore, MRTD is used to reduce the over-smoothness in the spectral sequence generated by HMMs in this research. Previous studies [10] have also found that event functions of MRTD can represent the "languageness" or content information of speech, which is important to perceive speech intelligibility, while event targets of MRTD can convey the non-linguistics of style information such as speaker individuality, which is important to perceive the naturalness of speech [10]. Therefore, the factors that are important to speech intelligibility such as the MRTD event functions of the spectral trajectory generated by HMM-based TTS are kept unchanged to preserve the high intelligibility of HMM-based TTS. Whereas the factors that are important to speech naturalness such as MRTD spectral event targets are rendered with an original speech database to enhance the naturalness of HMM-based TTS.

In the first stage of the proposed HTD, HMM-based TTS is used to generate spectral and prosodic trajectories. Previous results [11] show that HMM-based TTS is efficient on prosodic modeling but needs improvements on spectral modeling, particularly on reducing over-smoothness in spectral features. Therefore, prosodic features generated by HMM-based TTS are preserved. To reduce the over-smoothness in the spectral sequence generated by HMM-based TTS and to transform it to be close with that of the original speech, this spectral sequence is analyzed by MRTD to obtain the corresponding event targets and event functions. These event functions are also preserved due to the relations between event functions of MRTD and the speech intelligibility and since speech synthesized by HMM-based TTS is already highly and stably intelligible. However, these event targets are rendered and replaced with closest neighbors in an original database to make the spectral sequence generated by HMM-based TTS to be transformed to that of the original speech. Then, over-smoothness in the spectral sequence generated by HMM-based TTS can be reduced and the detail information in the spectral sequence of the original speech, related to the perception of the naturalness of speech, can be recovered. The smoothness of event functions of MRTD ensures the spectral sequence is still smooth after being rendered even with a limited amount of data for rendering. As a result, speech synthesized by the proposed HTD can be not only highly intelligible but also natural, even when the size of the database for rendering is limited. Additionally, the footprint size of the proposed HTD can be small due to the sparse representation of MRTD.

2. The Proposed HTD

2.1 Outline of HTD

There is a diagram of the proposed HTD in Fig. 1.

The spectral and prosodic trajectories are generated from HMM-based TTS in the first stage. Since HMM-based TTS is efficient on prosodic modeling [11], the prosodic trajectories of the F0 contour and gain contour of HMM-based TTS are preserved for the proposed HTD.

A sequence of the line spectral frequency (LSF) or the line spectral pairs (LSP) generated by the HMM-based TTS is analyzed in the second stage by MRTD [9]. Assume that $\boldsymbol{y(n)}$ is this spectral sequence, MRTD decomposes $\boldsymbol{y(n)}$ into K dynamic event functions ϕ_k and K static event targets $\boldsymbol{a_k}$ and $k = 1..K$, as given in Eq. (1). Here, $\hat{\boldsymbol{y}}\boldsymbol{(n)}$ is the approximation of $\boldsymbol{y(n)}$. There are K event targets in a total of

N frames and $K << N$, then MRTD (and TD in general) is a sparse representation of speech. The event functions are interpolation functions representing temporal transition movements between the sparse event targets.

$$\hat{\boldsymbol{y}}(\boldsymbol{n}) = \sum_{k=1}^{K} \boldsymbol{a_k}\phi_k(n), 1 \le n \le N \tag{1}$$

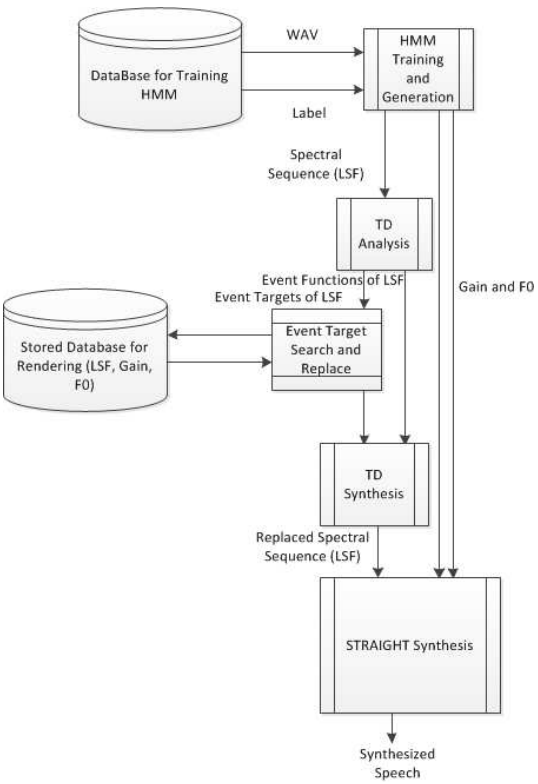

Fig. 1 Overview of HTD.

Equation (1) can be written in matrix notation as Eq. (2), where P is the dimension of the speech parameter.

$$\hat{Y}_{P\times N} = \boldsymbol{A_{P\times K}}\Phi_{K\times N} \tag{2}$$

Figure 2 draws an example of MRTD with spectral parameter $y(1 : N)$, event targets $a(1 : K)$, and event functions $\phi(1 : K)$.

Event target $\boldsymbol{a}$ and event function ϕ are unknown in Eqs. (1) and (2) and need to be estimated by using some optimization tasks to minimize interpolation error.

In the first step of the optimization task in MRTD [9], event targets are set equal to the frame-based vector at the same locations as given in Eq. (3).

$$\boldsymbol{a_k} = y(n_k) \tag{3}$$

Here, n_k is the location of event target a_k.

In the second step of the optimization task, event functions in MRTD are estimated as described in Eqs. (4) and (5). Here, $\langle \ldots \rangle$ and $\|.\|$ correspond to the inner product of two vectors and the norm of a vector.

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), \; if \; n_{k-1} < n < n_k \\ 1, \; if \; n = n_k \\ \min(\phi_k(n-1), \max(0, \hat{\phi}_k(n))), \\ if \; n_k < n < n_{k+1} \\ 0, \; otherwise \end{cases} \tag{4}$$

$$\hat{\phi}_k(n) = \frac{\langle (y(n) - a_{k+1}), (a_k - a_{k+1}) \rangle}{\|a_k - a_{k+1}\|^2} \tag{5}$$

Using the estimation given in Eqs. (4) and (5), each event function $\phi_k(n)$ is smooth, has only one peak, and two overlapped event functions sum up to one as described in Fig. 2 and explained in detail in [9]. These properties of event functions results in gradual movements of the interpolated spectral $\hat{\boldsymbol{y}}(\boldsymbol{n})$ that are related to the co-articulation of speech.

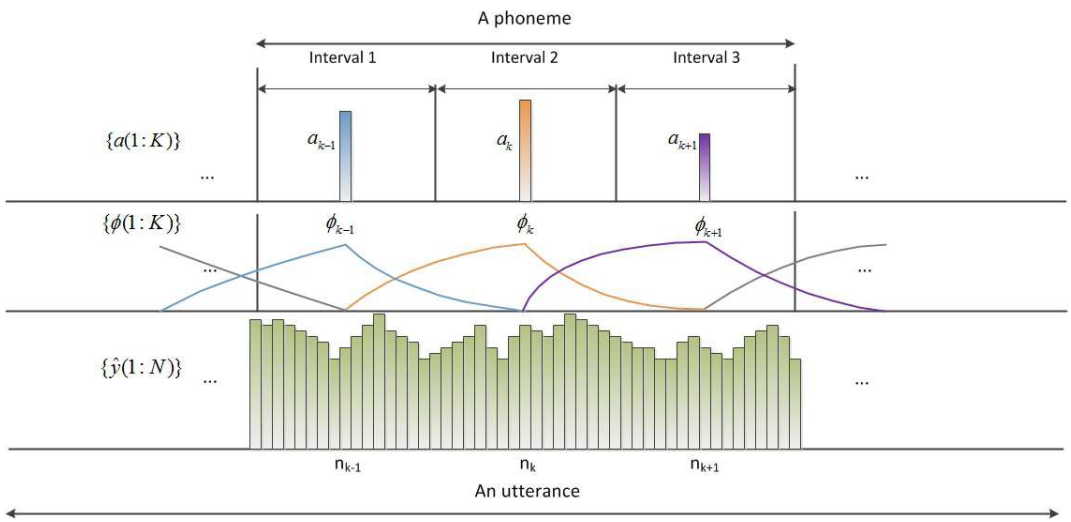

Fig. 2 An example of MRTD analysis / synthesis with N frames and K event targets: a bar represents a frame-based spectral feature or a spectral event target in a specific location. Two bars located at two locations of the same spectral event target are same in lengths.

In addition, the modification on sparse event targets $\boldsymbol{a_k}$ directly and gradually affects to all frames inside duration in which the event function ϕ_k is non-zero. Hence, speech can be flexibly modified / transformed at specific events in the time domain by modifying / transforming MRTD event targets $\boldsymbol{a}$ as shown in [10].

After the event functions are estimated, the event targets are re-estimated in the last step of the optimization task as shown in Eq. (6) to minimize the interpolation error, where ^{T} is matrix transpose transformation.

$$A = Y\Phi^T(\Phi\Phi^T)^{-1} \tag{6}$$

Note that Eq. (6) is the general form of the re-estimation of event targets of LSF in the original MRTD that was described in details in the original work [9]. For short, Eq. (6) means that each event target is re-estimated by its initialized value, which is the frame-based vector at the same location, and the non-zero estimated event functions at the same location with a convergence condition of minimizing the reconstruction error and ensuring the orders of LSF.

The event functions of the spectral sequence generated from HMM-based TTS, which are important to perceive speech intelligibility, are preserved in the third stage of the proposed HTD, because HMM-based TTS is already stable and highly intelligible. The target vectors are modified by selection from an original dataset to overcome oversmoothness in the spectral sequence generated by HMM and to render the spectral sequence close to that of original speech. The procedure for target selection is described in more detail in the next sub-section.

Finally, the high-quality speech vocoder STRAIGHT [12] is used to generate speech waveforms.

2.2 Target Selection Procedure

As the proposed method for selection is based on event targets, the concept behind the proposed selection procedure can be considered to be a new concept of "target selection" rather than the conventional concepts "unit selection" and "frame selection" [7].

The event targets of the speech trajectory generated by HMM are modified in the proposed HTD by replacing them with the most-matched event targets of original speech. Therefore, an alignment procedure in the time domain is required.

Dynamic time wrapping (DTW) or the nearest neighbor search (NNS) can be used in the frame-based voice transformation to align the transformation in parallel form for the former and in non-parallel form for the latter. A technique of using a fixed number of equally-spaced event targets for each phoneme has been shown to be flexible and efficient for TD-based voice transformations [10]. This method involves non-parallel transformation for a syllables or an utterance but is a parallel transformation for each phoneme when each ordered event target of a source phoneme is transformed into a corresponding ordered event target of a target phoneme. Developing from this method, each phoneme is divided into three equally-spaced intervals in this work. One event target is located at the center of each of the three intervals. Therefore, there are three event targets in one phoneme. The number of event targets in one phoneme can be from one as in the original MRTD [9], or five in [10]. There are two reasons for choosing three event targets in one phoneme in this work. The first one is that increasing the number of event targets in one phoneme larger than three does not improve the quality of synthesized speech in our experiments, but increases the size of stored data for rendering. The second one is that we want to set the number of equally-spaced intervals as well as the number of event targets in one phoneme same as the number of HMM states in each phoneme, which is three in this work, with an expectation that all HMM states are rendered by the original data. Although the method of locating event targets at center frames in each HMM state in Viterbi alignment is straightforward and may increase the accuracy of the selection procedure, this method has not implemented in this research at present. This is one of our future works.

The event targets are searched and replaced as described in Fig. 3. Each event target of the source spectral sequence generated by HMM is replaced by an event target of the original speech.

Using MRTD analysis, each event target is re-estimated by the frame-based vector at the same location, and the es-

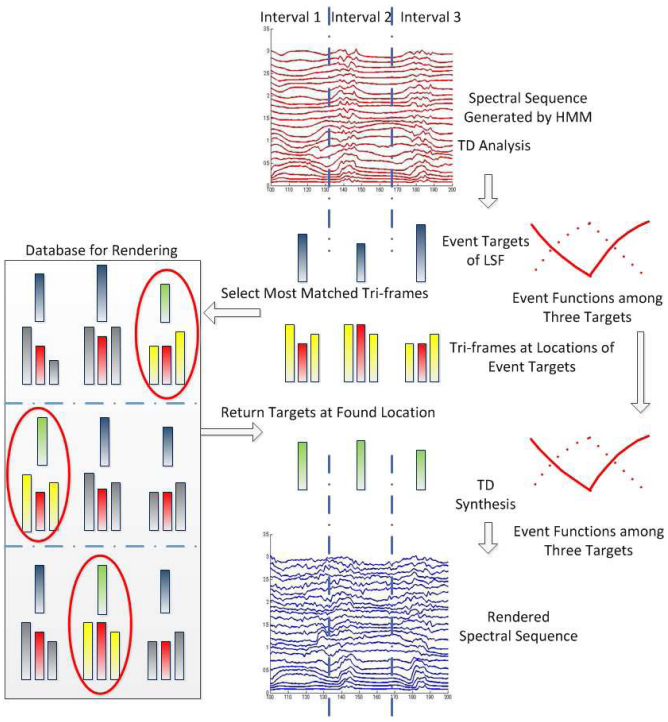

Fig. 3 Target Selection: Single bars represent spectral event targets located at centers of equally-spaced intervals; triple bars represent frame-based features in tri-frames where their central frames are located at the same positions as event targets. Input synthetic tri-frames color yellow-red-yellow, selected original tri-frames with the same colors are marked by red circles, and green event targets are the event targets of the original speech selected for replacing.

timated event function at the same location, as explained in sub-section 1. Therefore, event targets depend on the wide-range context and sensitive to its locations. As a result, to directly use event targets for alignment may reduce the accuracy of the alignment procedure. Instead of that, three consecutive frames, referred to as tri-frames in this research, located at same positions of event targets, are used to align the source and target event target pairs.

The matched tri-frames of the source - target pairs $s-t$ are searched by NNS with a summed cost as defined in Eqs. (7), (8), (9), (10), and (11) with the sub-costs of fundamental frequency (F0), LSF with order P, and power gain (PL).

$$d = N(d_{F0}) + N(d_{LSF}) + N(d_{PL}) \tag{7}$$

$$d_{F0} = \left|\log(F0_t) - \log(F0_s)\right| \tag{8}$$

$$d_{LSF} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(LSF_{i,t} - LSF_{i,s})^2} \tag{9}$$

$$d_{PL} = \left|\log(PL_t) - \log(PL_s)\right| \tag{10}$$

$$N(d) = \frac{d - \mu_d}{\sigma_d} \tag{11}$$

Each component cost is normalized by normal distribution similarly to that with HTT to avoid of weighting the component costs [7], as shown in Eq. (11), where μ_d and σ_d correspond to the mean and standard deviation of the sample distances of all candidates.

In our implementation, the "target selection" is supervised by label data to ensure its accuracy and reduce the length of searching time, in which each ordered target in a phoneme is replaced by the selected targets with the same order and in the same phoneme.

In the offline stage, the database for rendering is prepared with two steps. First, all utterances with labels are analyzed by MRTD. Then, analyzed event targets and tri-frames at the same locations are extracted from the parameters of the whole utterances by using label data, and stored for each distinct phoneme.

In the online rendering stage for each phoneme, the matched original tri-frames are selected from the original data and the event targets of the spectral sequence generated by HMM-based TTS in the previous stage are replaced by the original event targets located in the same positions as the selected original tri-frames. The "target selection" will be run with the whole database if the target phoneme for rendering is not found. Therefore, the selection procedure can still work if the number of phonemes in the database for rendering is not sufficient, such as under some limited data conditions, for instance.

2.3 Differences between HTD and HTT

Although the proposed HTD shares some common procedures with HTT, their concepts are completely different. These differences are presented and discussed in this section. There are four main differences:

(1) HTT can be considered to be one kind of unit selection that uses HMM-based TTS to compute the target cost, resulting in improved stability in synthesized trajectory of speech. However, HTT shares several common disadvantages with unit selection TTS, e.g., their requirements for huge amounts of data for selection or rendering, their huge footprints, and their inflexibility for voice transformations. The proposed HTD is one kind of HMM-based TTS that uses MRTD and a "target selection" procedure to reduce over-smoothness, resulting in the improvement of synthesized speech in terms of naturalness while other advantages of HMM-based TTS can be still preserved.

(2) HTT requires a huge database for rendering to ensure the smoothness of the synthesized speech since limited data may cause mismatches and discontinuities between consecutive frames. The smoothness of the synthesized trajectory in HTD is ensured by the smoothness of event functions and the stability and smoothness of the trajectory generated by HMM-based TTS. Therefore, the matching level of the "target selection" task does not strictly require precision as in HTT. As a result, HTD can synthesize stable and smooth speech even under limited data conditions.

(3) HTT has a large footprint because it requires a huge database for "frame selection". HTD can have a small footprint because the sparse "target selection" can be used with small databases for rendering. Even when the same database is used for rendering, the sparse "target selection" also stores a smaller footprint compared with the "frame selection" in HTT.

(4) HTT can be combined with voice transformation by using multiple huge target databases for rendering [7]. The requirement for huge target databases is not convenient for practical voice transformations where only a few target data are available. TD-based voice transformations [10] could efficiently transform speaker individuality by preserving the event functions of source speech and transforming its event targets to those of target speech. This manner is similar to the proposed HTD, when event functions of speech synthesized by HMM-based TTS are preserved and its event targets are selected from an original database. As aforementioned, the "target selection" does not require a huge database. Therefore, HTD can be flexibly combined with voice transformation by using multiple small target databases for rendering.

3. Performance Evaluations

3.1 Data Preparation

TTS under limited data conditions is more practical for under-resourced languages, where huge public speech corpora are missing, compared with highly-resourced languages. Vietnamese is a language spoken by about 100 million people throughout the world. However, as there is no huge public speech corpus with labeling for Vietnamese at

present, it is an under-resourced language.

Vietnamese is a tonal monosyllabic language. There are about 7000 distinct Vietnamese tonal syllables. There are totally 20 consonants and 250 tonal vowels in Vietnamese. More detail on Vietnamese can be found in [13]. In this research, we used the small Vietnamese corpus DEMEN567, including 567 utterances. This corpus was also called TTSCorpus in [14]. The total time interval of this dataset is approximately one hour. The sampling frequency of the corpus is 11025 Hz.

The objective for this research is to propose an efficient TTS under limited data conditions. In this research, a dataset in this study is considered to be under "limited data conditions" if it reaches a threshold when the phoneme coverage is approximately 100 %. All phonemes exist but their frequencies are small due to this requirement. Therefore, the estimation of HMM for one phoneme may be inappropriate since there is small amount of training data for this phoneme. Over-smoothness in HMM-based TTS is significant in this case and improvements to the proposed HTD are more important. A "limited data condition" was simulated by taking this requirement into account with a dataset of 300 utterances extracted from DEMEN567. This dataset is close to the threshold where the phoneme coverage reaches approximately 100%. Although some monophones are still missing, most of widely used tonal phonemes appear in this dataset. The size of this dataset in PCM 16 bits format is approximately 30MBs and the duration is approximately 20 minutes. This dataset was used to train the HMM-based TTS, which was used as input of HTT and the proposed HTD, and was used for comparisons.

We used three datasets including 100, 300, and 500 utterances, extracted from DEMEN567, for rendering HTT and the proposed HTD to investigate the dependence of the performances of HTT and HTD on the sizes of the databases used for rendering. Note that HMM-based TTS was just trained one time with the dataset of 300 utterances, simulated to be an "under limited data condition".

3.2 Experimental Parameters

We compared five versions of speech in our evaluations: speech synthesized by a HMM-based TTS for Vietnamese [15] trained with 300 utterances, speech synthesized by HTT, speech synthesized by the proposed HTD, speech analyzed / synthesized by MRTD-STRAIGHT, and the original speech. HTT and our proposed HTD used 100, 300, and 500 utterances for rendering. Speech analyzed / synthesized by MRTD and STRAIGHT can be considered as the ideal limitation of HTD obtained when using a huge amount of data for rendering. Due to reconstruction errors of MRTD and STRAIGHT, this ideal limitation of HTD is different from the original speech. The original speech can be considered as the ideal limitation of HTT when using a huge amount of data for rendering since HTT is one kind of waveform concatenation TTS used the original speech. Although these two ideal limitations of HTD and HTT can be never reached, they were used for evaluations in this paper instead of evaluating HTD and HTT with a real large-scaled speech corpus because the latter solution is expensive, time-consuming, and not available for us at present.

All experimental parameters were controlled to be equivalent for all TTSs to enable them to be fairly evaluated. The spectral features for the TTSs were LSF with an order of 24. The HMM-based TTS also used the deltas of LSF. The excitation parameters for HMM-based TTS were composed of logarithmic F0 and their corresponding delta coefficients. The frame length was 20-ms and the update interval was 5-ms. The context-dependent HMM used three states for one phoneme, which was same as the number of event targets for one phoneme that was used in the proposed HTD. Other parameters of the HMM-based TTS for Vietnamese were adopted from the original work by Vu et al. [15], while those of HTT were adopted from the original work by Qian et al. [7].

STRAIGHT version 4 [12] was used as a vocoder to generate the output waveforms. All parameters used for extracting F0, aperiodicity (AP), and spectral envelope with STRAIGHT were default parameters except for f_s, frame size and frame step.

3.3 Subjective Evaluations

Subjective tests on intelligibility and naturalness were conducted to evaluate the TTSs. Five subjects who were native Vietnamese with normal hearing participated in these tests.

Semantically unpredictable sentences (SUSs) have been used as a standard measure to evaluate the intelligibility of a TTS, but there are no designs on Vietnamese SUS lists at present. Therefore, 20 testing sentences were chosen with four restricted rules (rules 1–4) to prevent the subjects from easily predicting the meanings, and two restricted rules (rules 5–6) were chosen to ensure the evaluations were reliable. The six rules were: (1) the Vietnamese words in the testing sentences were all low frequency, (2) only sentences composed of monosyllabic words were used to prevent subjects from predicting the meaning of compound words with only their constituent parts, (3) repeating the words between testing sentences was avoided to prevent subjects from remembering words that they had heard previously, (4) sentences with fewer semantic relations were selected to prevent subjects from predicting the meaning of sentences, (5) sentences covering all Vietnamese tones and minimizing the repetition of tonal phonemes were selected, and (6) only short sentences were selected to avoid the difficulty for subjects to remember the syllables that they had heard in the testing sentences. In this research, the intelligibility scores were measured by word error rates (WER) of SUS sentences.

The naturalness of TTS has been widely evaluated by mean opinion scores (MOS). Therefore, MOS scores were used to evaluate the overall impression of naturalness of TTSs with a testing dataset that contained 20 long sentences with an average length approximately 25 syllables. Evalua-

tions with long sentences were used to measure the speech naturalness in terms of both voice quality and segmental duration and timing.

The two testing datasets for intelligibility and naturalness evaluations were chosen from the set of sentences that were not used for training the HMM-based TTS and were not used for rendering with HTT and the proposed HTD.

3.3.1 Results of the Intelligibility Evaluation

The results obtained from the intelligibility evaluations are listed in Table 1. They indicate that the WERs of original speech are zeros and those of speech synthesized by HMM-based TTS were small and were equivalent to those of speech analyzed / synthesized by MRTD-STRAIGHT. Both HTT and HTD reduced the intelligibility of speech synthesized by HMM-based TTS. A statistical F-test was conducted to investigate how much HTT and HTD reduced the intelligibility in three conditions for rendering. The results are given in Tables 2 and 3, which indicate that HTT significantly reduced the intelligibility of HMM-based TTS while the reduction with HTD was not significant. With 500 utterances for rendering, the intelligibility of HTD was even equivalent with that of HMM-based TTS and of MRTD-STRAIGHT. As a consequence, the proposed HTD was successful to preserve the intelligibility of HMM-based TTS under limited data conditions.

3.3.2 Results of the Naturalness Evaluation

The results from the naturalness evaluations are presented in Fig. 4, in which the MOS scores of speech synthesized by HMM-based TTS, analyzed / synthesized by MRTD-STRAIGHT, and the original speech were drawn as the same values in the three conditions, for convenience to compare all conditions for rendering. These results indicate that HTT improved the naturalness of HMM-based TTS trained under limited data conditions with 300 sentences when using a sufficient amount of data for rendering, i.e 300 and 500 sentences. However, HTT reduced the naturalness of HMM-based TTS when using 100 utterances for rendering. These results also indicate that HTD improved the naturalness of HMM-based TTS trained under limited data conditions with 300 sentences when using all three datasets for rendering.

A statistical F-test was conducted to measure the significance of the improvements and reductions on naturalness of HTT and HTD compared with the HMM-based TTS trained with 300 sentences. The results are shown in Tables 4 and 5. They indicate that HTT significantly reduced the naturalness of the HMM-based TTS when using 100 utterances for rendering and insignificantly improved the naturalness of the HMM-based TTS when using 300 and 500 utterances for rendering. They also indicate that HTD significantly improved the naturalness of HMM-based TTS in all three datasets used for rendering.

Consequently, the proposed HTD demonstrated its efficiency compared with HMM-based TTS and HTT in terms of naturalness in all conditions for rendering. Especially, the proposed HTD could improve the naturalness of HMM-based TTS even when using an ultra-small dataset for rendering, i.e. a dataset of 100 utterances.

Table 1 Means of WERs (%): HMM-based TTS was only trained with 300 utterances, speech analyzed / synthesized by MRTD-STRAIGHT and the original speech were independent with the datasets for rendering.

	100 Utterances	300 Utterances	500 Utterances
Original	-	0	-
MRTD-STRAIGHT	-	0.25	-
HMM-based TTS	-	0.25	-
HTD	0.64	0.51	0.25
HTT	7.13	3.82	3.69

Table 2 F-test to show differences in the intelligibility of HMM-based TTS and HTD in three conditions for rendering: No difference when using 500 utterances for rendering HTD.

	100 Utterances	300 Utterances	500 Utterances
F	21.008	17.049	-
p	< 0.001	< 0.001	-

Table 3 F-test to show differences in the intelligibility of HMM-based TTS and HTT in three conditions for rendering.

	100 Utterances	300 Utterances	500 Utterances
F	523.213	151.545	234.607
p	< 0.001	< 0.001	< 0.001

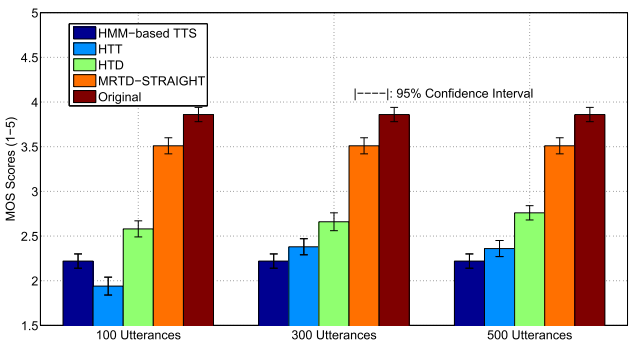

Fig. 4 Mean MOSs for naturalness evaluations and their 95% confidence intervals in three conditions for rendering HTT and HTD.

Table 4 F-test to show differences in the naturalness of HMM-based TTS and HTT in three conditions for rendering.

	100 Utterances	300 Utterances	500 Utterances
F	18.751	4.918	6.424
p	< 0.001	0.028	0.012

Table 5 F-test to show differences in the naturalness of HMM-based TTS and HTD in three conditions for rendering.

	100 Utterances	300 Utterances	500 Utterances
F	31.749	45.213	84.644
p	< 0.001	< 0.001	< 0.001

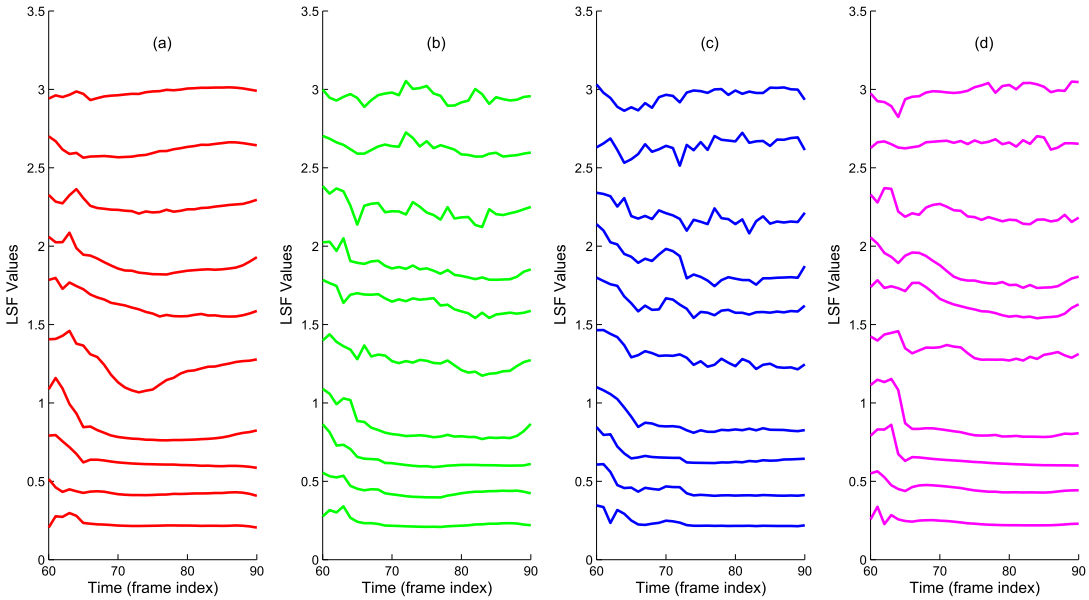

Fig. 5 LSF sequences: (a) synthesized by a HMM-based TTS trained with 300 utterances, (b) synthesized by a HTD rendered with 300 utterances, (c) synthesized by a HTT rendered with 300 utterances, (d) of the original speech.

3.4 Discussions

3.4.1 Discussions on Intelligibility and Naturalness

One sample of an LSF sequence synthesized by HMM-based TTS, HTD, HTT, and one of the original speech are given in Fig. 5. It reveals that both HTD and HTT can sharpen the LSF sequence generated by HMM-based TTS. However, the frame-based method in HTT may have excessive sharpening and may increase the discontinuities between frames under limited data conditions, resulting in decreased intelligibility and naturalness of HTT under limited data conditions.

Results from evaluations demonstrated that the proposed HTD with a new rendering method preserved the intelligibility and improved the naturalness of HMM-based TTS by reducing the problem with over-smoothness. The outperformance of the proposed HTD compared with the HTT confirmed that the proposed HTD is specifically efficient under limited data conditions. However, the results from intelligibility and naturalness evaluations of MRTD-STRAIGHT, which is the ideal limitation of HTD, and the original speech show that HTT can be superior to HTD if a huge amount of data is used for rendering HTT and HTD.

The results from the intelligibility evaluation were consistent with the results from HMM-based TTS [15] where the intelligibility scores of a Vietnamese TTS could reach 100%. The intelligibility of the mono-syllabic Vietnamese speech seems to be higher than that of other languages. MOS score of 3.8 for the original speech was quite low since corpus DEMEN567 was not well recorded due to a low sampling frequency of 11025 Hz and the recording environment. The MOS scores for all HMM-based TTS, HTT and the proposed HTD were not high compared with those of the original speech since they were implemented under a "limited data condition".

3.4.2 Discussions on the Footprint Size

The footprint size of the proposed HTD is the sum of the size of the trained HMM model's parameters and the size of the stored data for rendering. The size of the trained HMM model's parameters is up to about $1-2MB$ if an appropriate fixed-point representation is used.

The experimental frame period was $5-ms$ and $f_s = 11.025$ KHz. If each sample is represented by N bytes with fixed-point representation, the size of each original frame period is $5 \times 11.025 \times N \approx 55 \times N$ bytes. The experimental event rate was approximately one target in eight frames on average by following the method of determining event locations presented in sub-section 2.2. Three parameters, 24-ordered LSF, F0, and PL, were used. Each event target of each of the three parameters was stored together with their corresponding tri-frames. If each value is represented by N bytes with fixed-point representation, the size of each encoded event target is $26\times(3+1)/8\times N = 13\times N$ bytes. Thus, the compression rate $rt \approx 55/13 \approx 4$. The sizes of the three original waveform databases of 100, 300, and 500 utterances for rendering were approximately 10 MBs, 30 MB, and 50 MBs, respectively. Therefore, the actual size of the smallest database for rendering was approximately $10/4 \approx 2.5$ MB, and the footprint of the proposed HTD was approximately $4-5$ MB. Although the proposed HTD increases the size of footprint compared with that of HMM-based TTS, it was still small enough for most limited-resourced hardware platforms. If further compression techniques are used such as vector quantization to quantize LSF, F0, and PL, the size of

the compressed footprint can be reduced even further.

3.5 Future Works

Due to time-consuming, only one single-speaker small Vietnamese speech corpus was used for evaluations. Therefore, the speaker-independence and language-independence of the proposed HTD were not investigated in this research. We plan to implement and to evaluate the proposed HTD with different multi-speaker databases in different languages to confirm whether this new approach is speaker-independent and language-independent.

The implementations of the proposed HTD still have remaining issues that should be improved. For example, event targets should be located at center frames in each HMM state in order to render speech synthesized by HMM-based TTS in all HMM states. Since it is possible to develop the proposed HTD to synthesize multiple voices with limited target data, voice transformations combined with the proposed HTD will be implemented and evaluated in the future.

4. Conclusion

The motivation for this paper is to improve the naturalness of HMM-based TTS trained under limited data conditions while preserving its intelligibility. To achieve this motivation, a hybrid TTS between HMM-based TTS and MRTD named HTD in this paper was proposed. Prosodic trajectories and event functions of the spectral trajectory generated by HMM-based TTS were kept unchanged to preserve the high intelligibility of HMM-based TTS. Whereas event targets of the spectral trajectory generated by HMM-based TTS were rendered with an original speech database to enhance the naturalness of HMM-based TTS. The experimental results under limited data conditions show that the intelligibility of speech synthesized by the proposed HTD was reduced insignificantly, whereas the naturalness of speech synthesized by the proposed HTD was improved significantly, compared with those by HMM-based TTS. Further discussions also show that HTD has small footprint. Therefore, the proposed HTD showed its strong efficiency under limited data conditions.

Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

References

[1] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," Proc. SSW, 2002.
[2] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," IEICE Trans. Fundamentals, vol.E86-A, no.8, pp.1956–1963, Aug. 2003.
[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, 2000.
[4] M. Zhang, J. Tao, H. Jia, and X. Wang, "Improving HMM Based speech synthesis by reducing over-smoothing problems," Proc. ISCSLP, pp.1–4, 2008.
[5] 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, May 2007.
[6] K. Hashimoto, S. Takaki, K. Oura, and K. Tokuda, "Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2011," Blizzard Challenge, 2011.
[7] Y. Qian, F.K. Soong, and Z. Yan, "A unified trajectory tiling approach to high quality speech rendering," IEEE Trans. Audio Speech Language Process., vol.21, no.2, pp.280–290, 2013.
[8] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP, pp.81–84, 1983.
[9] PC. Nguyen, T. Ochi, and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.397–405, March 2003.
[10] P.N. Binh and M. Akagi, "Efficient modeling of temporal structure of speech for applications in voice transformation," Proc. Interspeech, pp.1631–1634, 2009.
[11] R.B. Chicote, J. Yamagishi, S. King, J.M. Montero, and J.M. Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," Speech Commun., vol.52, pp.394–404, 2010.
[12] H. Kawahara, "STRAIGHT, exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoust. Sci & Tech., vol.27, no.6, pp.349–353, 2006.
[13] H. Phe, Chinh ta Tieng Viet (Vietnamese Grammar), pp.9–15, Da Nang Publisher, 2003.
[14] L.C. Mai and D.N. Duc, "Design of Vietnamese speech corpus and current status," Proc. ISCSLP, pp.748–758, 2006.
[15] TT. Vu, MC. Luong and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," Proc. O-COCOSDA, pp.116–121, 2009.

Trung-Nghia Phung received his B.E. in electronic & telecommunication engineering from the Hanoi University of Technology (HUT) in 2002 and his M.E. in electronic & telecommunication engineering from the Vietnam National University Hanoi, in 2007. He has been a Ph.D. candidate at the School of Information Science of the Japan Advanced Institute of Science and Technology (JAIST) since 2009.

Thanh-Son Phan received his B.E. degree in computer science from HUT in 2000, and his M.S. degree in computer science from the Le Quy Don Technical University (LQDTU) in 2006. He is currently a Ph.D. student at the IoIT.

Thang Tat Vu received his B.E. and M.E in electronic & telecommunication engineering from HUT in 2002 and 2004, and his PhD from JAIST in 2008. He is currently a researcher at IoIT. He is a member of the Research Institute of Signal Processing (RISP).

Mai Chi Luong received her B.E. from the Faculty of Applied Mathematics of Kishnov University (former Soviet Union) in 1981, and her Ph.D. from IoIT in 1991. She has been working as a senior researcher of IoIT from 1982 to present and became an Associate Professsor in 2005. Her research interests include pattern recognition, machine learning, and speech recognition and synthesis. She is a member of the IEEE. Dr. Luong received the Kovalevskaia Award for her outstanding contributions to R & D in Vietnam in 2010.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. in Engineering from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. He worked at the Auditory and Visual Perception Research Laboratories (ATR) from 1986 to 1990. He has been on the faculty of the School of Information Science of JAIST since 1992 and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in human beings, and signal processing of speech. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the ASJ, the IEEE, the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA).