A Hybrid TTS between Unit Selection and HMM-based TTS under limited data conditions

Trung-Nghia Phung[1,2], *Chi Mai Luong*[2], *Masato Akagi*[1]

[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2]Institute of Information Technology, Hanoi, Vietnam

ptnghia@jaist.ac.jp, lcmai@ioit.ac.vn, akagi@jaist.ac.jp

Abstract

The intelligibility of HMM-based TTS can reach that of the original speech. However, HMM-based TTS is far from natural. On the contrary, unit selection TTS is the most-natural sounding TTS currently. However, its intelligibility and naturalness on segmental duration and timing are not stable. Additionally, unit selection needs to store a huge amount of data for concatenation. Recently, hybrid approaches between these two TTS, i.e. the HMM trajectory tiling TTS (HTT), have been studied to take advantages of both unit selection and HMM-based TTS. However, such methods still require a huge amount of data for rendering. In this paper, a hybrid TTS among unit selection, HMM-based TTS, and the Modified Restricted Temporal Decomposition (MRTD), named HTD, is proposed motivating to take advantages of both unit selection and HMM-based TTS under limited data conditions. Here, TD is a sparse representation of speech that decomposes a spectral or prosodic sequence into two mutually independent components: static event targets and correspondent dynamic event functions, and MRTD is a compact but efficient version of TD. Previous studies show that the dynamic event functions of MRTD are related to the perception of speech intelligibility, one core linguistic or content information, while the static event targets of MRTD convey non-linguistic or style information. Therefore, by borrowing the concepts of unit selection to render the event targets of the spectral sequence, and directly borrowing the prosodic sequences and the dynamic event functions of the spectral sequence generated by HMM-based TTS, the naturalness and the intelligibility of the proposed HTD can reach the naturalness of unit selection, and the intelligibility of HMM-based TTS, respectively. Due to the smoothness of event functions of MRTD, an appropriate smoothness in synthesized speech can still be ensured when being rendering by a small amount of data, resulting in the usability of the proposed HTD under limited data conditions. The experimental results with a small Vietnamese dataset, simulated to be a "limited data condition", show that the proposed HTD outperformed all HMM-based TTS, unit selection, HTT under a limited data condition.

Index Terms: TTS, unit selection, HMM-based, Temporal Decomposition, HTT

1. Introduction

Building a large-scale speech corpus is a costly task that takes a long time and a great deal effort by engineers, acousticians and linguists. Therefore, to build high-quality speech synthesizers under limited data conditions is important in practice, specifically for under-resourced languages.

The two most successful TTS up to now are unit selection [1] and HMM-based [2, 3]. HMM-based TTS provides synthetic speech with stable and smooth trajectory. Therefore, the intelligibility of HMM-based TTS can reach that of the original speech. However, the naturalness of HMM-based TTS is low mainly due to the spectral over-smoothness caused by the "averagely" statistical processing. Although much research has been attempted to reduce this over-smoothness [3], speech synthesized by HMM-based TTS is still muffled and far from natural. On the contrary, the naturalness of unit-selection TTS is high. However, unit-selection has some drawbacks reducing the range of its practical applications. Among them, one main drawback is the unstability of the temporal trajectory of speech synthesized by unit selection. This artifact reduces the intelligibility of the speech synthesized by unit selection TTS compared with that of HMM-based TTS [4]. Another main drawback of unit selection is the requirement of a huge data corpus for concatenation.

Recently, hybrid approaches, which use HMM to guide the unit selection process to improve the stability and the smoothness of unit selection TTS, have been shown as the most successful TTS that can preserve the advantages of both HMM-based and unit selection TTS. Among them, the HTT [4] can be considered as the state-of-the-art hybrid TTS. In this system, HMM trajectory is used to guide the selection of each 5ms frame to concatenate the waveforms. The naturalness of this hybrid TTS is comparative with that of state-of-the-art unit selection TTS, while the intelligibility of this hybrid TTS is comparative as that of state-of-the-art HMM-based TTS. However, HTT [4] still requires a huge amount of data for rendering due to the use of "frame selection". If the selection process is imperfect due to some reasons such as the limitation of the data corpus, it is easy to perceive the discontinuities between frames. The experimental results in [4] show that the quality of HTT is stable only if the amount of stored database for rendering is approximately 2 hours to 10 hours.

In this study, in order to borrow the high naturalness of unit selection and the high intelligibility of HMM-based TTS, MRTD [6] is used to decompose the spectral sequence generated by HMM into two mutually independent components: static event targets and correspondent dynamic event functions, in which one component is related to the perception of intelligibility and the other one is related to the perception of naturalness. These two components are independently controlled in order to reach the intelligibility of HMM-based TTS and the naturalness of unit selection TTS. The proposed method is supported by previous studies [7, 8], where the temporal event functions can represent the "languageness" or linguistic information of speech, which is important for the perception of speech intelligibility, while the sparse event targets can convey non-linguistic

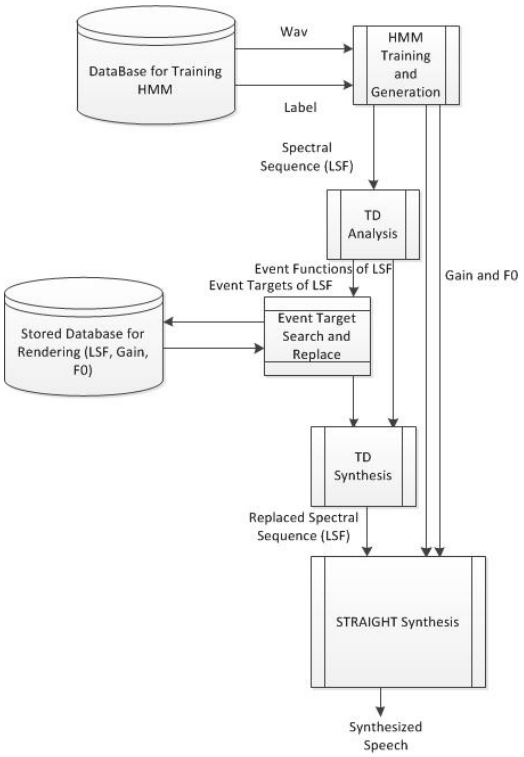

Figure 1: General Diagram.

of style information, which is important for the perception of the speech naturalness [7, 8].

Comparing with HTT [4], the proposed HTD has one main advantage on reducing the required amount of data for rendering, resulting in improved quality of the synthesized speech under limited data conditions. It reveals that HTT replaces all frames of the guided trajectory generated by HMMs with the closest frames found in the original database. Therefore, to ensure the stability and the smoothness of the synthesized speech, HTT requires a huge database for rendering because the limitation of data may cause the mismatches and the discontinuities between the consecutive replaced frames. However, in the proposed HTD, the smoothness of the spectral and prosodic trajectories is ensured by the smoothness of the event functions of the spectral sequence generated by HMM and the smoothness of prosodic sequences generated by HMM. Therefore, the matching level of the "target selection" task is not strictly required as in HTT [4], resulting in the reduction of the required amount of data for rendering.

In the next section, we will present and explain the details of the proposed HTD. Section 3 describes the evaluations, and finally section 4 draws the conclusions.

2. The proposed HTD

2.1. Outline of the proposed TTS

The general diagram of the proposed HTD is shown in Fig. 1.

At the first stage, spectral and prosodic trajectories are generated from HMM-based TTS. Since HMM-based TTS is efficient on prosodic modeling [5], the prosodic trajectories of the F0 contour and gain contour of HMM-based TTS are preserved for the proposed HTD.

At the second stage, the line spectral frequency (LSF) sequence generated by HMM-based TTS is analyzed by TD analysis [9] using a simplified TD version called MRTD [6].

Assume that the $\boldsymbol{y(n)}$ is this spectral sequence, TD decomposes $\boldsymbol{y(n)}$ into dynamic event functions ϕ and K static event targets $\boldsymbol{a}$ among total N frames, as given in Eqs. 1 and 2. Here, $\boldsymbol{\hat{y}(n)}$ is the approximation of $\boldsymbol{y(n)}$. There are K event targets in a total of N frames and $K << N$, then TD is a sparse representation of speech. The event functions are interpolation functions representing temporal transition movements between the sparse event targets. [9].

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

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

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

In Eq. (1) (or matrix representation as in Eq. (2)), event target $\boldsymbol{a}$ and event function ϕ are unknown and needed to be estimated by some optimization tasks to minimize interpolation error. In the initiation of the optimization task in MRTD [6], 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.

Then, event functions in MRTD are estimated as described in Eqs. (4) and (5). Here, $\langle ... \rangle$ 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 [6]. These properties of event functions results in gradual movements of the interpolated spectral $\boldsymbol{\hat{y}(n)}$ that are related to the co-articulation of speech. 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 [7, 8].

After estimating event functions, event targets are estimated as in Eq. 6, 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 [6]. 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

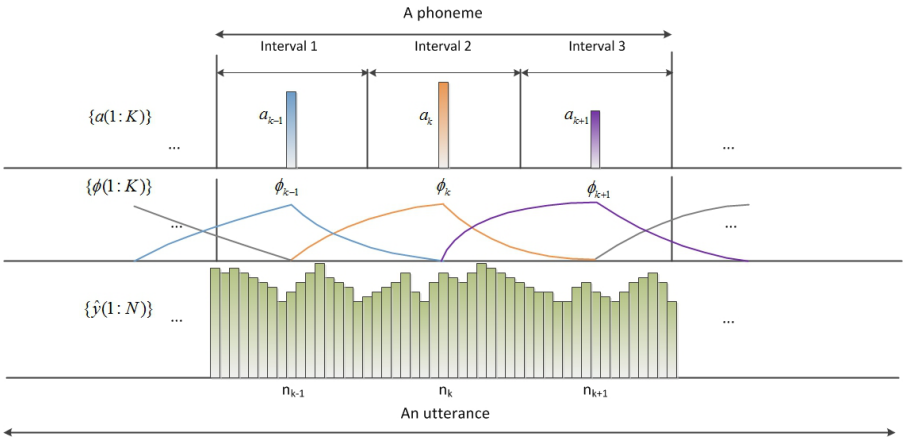

Figure 2: An example of MRTD analysis / synthesis with N frames and K event targets: a bar represents a frame-based spectral feature or a spectral event target in a specific location. Two bars located at two locations of the same spectral event target are same in lengths.

condition of minimizing the reconstruction error and ensuring the orders of LSF.

In the third stage of the proposed HTD, the event functions ϕ_k of the spectral sequence $\boldsymbol{y}$, which are important for speech intelligibility as mentioned in section 1, are also directly borrowed from HMM-based TTS because the intelligibility of HMM-based TTS is stable and high, approximately that of original speech. To overcome the over-smoothness of the spectral sequence generated by HMM and also to transform the spectral sequence to the original speech, the event targets $\boldsymbol{a_k}$ are selected from an original dataset. The target selection procedure is described more detail in the next sub-section.

Finally, at the last stage, the high-quality speech vocoder STRAIGHT [10] is used to generate speech waveforms.

2.2. Event target selection

The proposed method for selection is based on event target, so the proposed TTS can be considered as a new concept "target selection" rather than conventional concepts "unit selection" and "frame selection" [4].

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 to accurately modify the event targets of source speech into the corresponding event targets of target speech.

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 proposed when using TD-based voice transformation [7]. This method involves non-parallel transformation for a syllable 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 [6], or five in [7]. 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, compared with the use of equally-spaced intervals in the proposed method, this method has not implemented in this research at present. This is one of our future works.

The event target searching and replacing are represented in Fig. 3. Each event target of source spectral sequence generated by HMM is replaced by an event target of the original speech, searched by a selection process. The selection is supervised by labeled data in order to ensure its accuracy and reduce the amount of searching time. Using MRTD analysis, each event target is re-estimated by the frame-based vector at the same location, and the estimated non-zero event functions at the same location, as explained in sub-section 2.1. 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 are searched by nearest neighbor search with a summed cost as defined in Eqs. 7, 8, 9, 10, and 11.

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

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

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

$$d_G = |\log(G_t) - \log(G_s)| \tag{10}$$

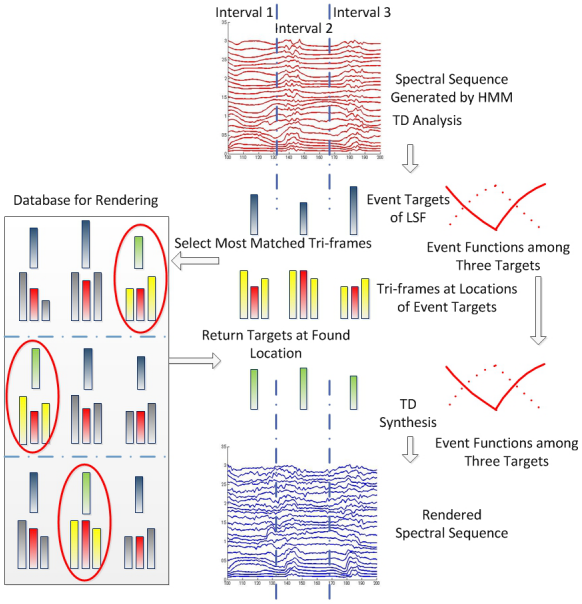

Figure 3: Target Selection: Single bars represent spectral event targets located at centers of equally-spaced intervals; triple bars represent frame-based features in tri-frames where their central frames are located at the same positions as event targets. Input synthetic tri-frames color yellow-red-yellow, 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.

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

Here, each cost for LSF, F0, and gain G is computed with a source target pair s and t. Each component cost is normalized by normal distribution similar as in [4], as shown in Eq. 11, where are mean and standard deviation of the sample distances of all candidates, respectively. The ideal behind the use of all F0, LSF, and PL to compute the distance cost between the trajectories generated by HMM-based TTS and those of the original speech in HTT is to find the physically closest frames (in the waveform domain) for concatenation. This ideal was adopted to the proposed HTD to select the spectral target of the physically closest frame in the original database.

After the matched tri-frames have been selected from the original data, the event targets of the spectral sequence generated by HMM-based TTS in the previous stage are replaced by the outputs of the selection procedure, which are the original event targets located in the same positions as the selected tri-frames.

In our implementation, the "target selection" is supervised by labeled 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.

3. Implementations and Evaluations

3.1. Data preperation

The main motivation of this study is to propose an efficient hybrid TTS under limited data conditions. The TTS under limited data conditions is more practical for under-resourced languages, where huge public speech corpus is missing, compared with high-resourced languages. Vietnamese is a language spoken by about 100 million people in the world. However, there is no huge public speech corpus with labeling for Vietnamese at present. Therefore, Vietnamese is one of the under-resourced languages and a Vietnamese corpus is used in this study.

Vietnamese is a tonal monosyllabic language. There are 20 consonants and 250 tonal vowels in Vietnamese. More detail on Vietnamese can be found in [11]. In this research, we used the small Vietnamese corpus DEMEN567, including 567 utterances. This corpus was also called TTSCorpus in [12]. The total time interval of this dataset is approximately one hour. The sampling frequency of the corpus is 11025 Hz. We simulated a limited data condition by a dataset of 300 utterances. This dataset is close to the threshold where the phoneme coverage reaches approximately 100%. Although some monophones are still mising, 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.

3.2. Experimental Parameters

We compared the six versions of speech in our evaluations: speech synthesized by a HMM-based TTS for Vietnamese [13], speech synthesized by a non-uniform unit selection TTS for Vietnamese [14], speech synthesized by HTT, speech synthesized by our proposed HTD, speech analyzed / synthesized by MRTD-STRAIGHT, and the original speech. All speech synthesizers used the same dataset simulated to be "under limited data conditions".

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 unit selection TTS and HTT when using a huge amount of data for selection or rendering since these synthesizers are waveform concatenation TTSs that use the original speech. Although these two ideal limitations can be never reached, they were used for evaluations in this paper instead of evaluating the synthesizers 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 three 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 loga-

rithmic F0 and their corresponding delta coefficients. The frame lengths ware 20 ms and the update intervals ware 5 ms. The context-dependent HMM used three states for one phoneme, which was the 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. [13], while those of HTT were adopted from the original work by Qian et al.[4] and those of unit selection TTS were adopted from the original work in[14].

STRAIGHT version 4 [10] 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 fs, frame size and frame step.

3.3. Subjective evaluations

For evaluating the TTS, subjective tests on intelligibility and naturalness were conducted. Five subjects who are native Vietnamese with normal hearing were required to attend the subjective tests.

Semantically unpredictable sentences (SUS) have been used as a standard measure to evaluate the intelligibility of a TTS, but there are no designs on Vietnamese SUS sentence lists at present. Therefore, a dataset of 20 testing sentences were chosen for the intelligibility evaluation with four restricted rules of preventing the subjects from predicting the meanings easily (rules 1–4), and two restricted rules for ensuring the reliability of the evaluation (rules 5–6):

(1) The Vietnamese words in the testing sentences were all low frequency;

(2) Only sentences composed from monosyllabic words were used to avoid ubjects from predicting the meaning of complex words with only their component words;

(3) Repeating the words between testing sentences is avoided, in order to prevent subjects remembering the words they heard previously;

(4) The sentences with less semantic relations were selected to avoid subjects predicting the meaning of the sentence;

(5) The sentences covering all Vietnamese tones and minimizing the repetition of tonal phonemes were selected;

(6) Only short sentences were selected to avoid the difficulty of subjects remembering the syllables that they heard in the testing sentence. 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. Evaluations 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 were chosen from the set of sentences that were not used for training, or concatenating, or rendering the TTSs.

The results of the intelligibility evaluations are shown in Table. 1. These results shows that the WERs of speech synthesized by HMM-based TTS, that of speech synthesized by the proposed HTD, and that of speech analyzed / synthesized by MRTD-STRAIGHT were very small. *The intelligibility of HMM-based TTS and HTD were equivalent and they highly outperformed the intelligibility of HTT and unit selection TTS.*

The results of the naturalness evaluation are shown in Fig. 4. *These results in terms of naturalness show that the proposed HTD was superior to all the HMM-based TTS, the unit selection TTS and the HTT.*

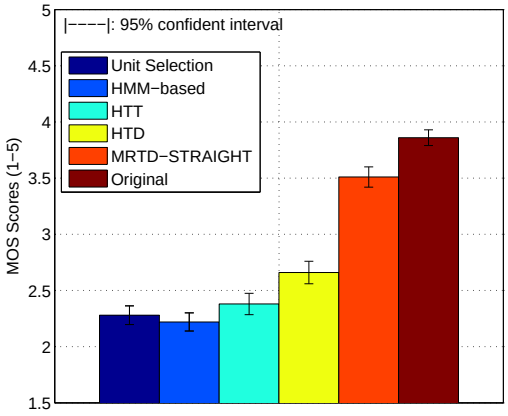

Figure 4: Naturalness mean scores and 95% confidence intervals for long sentences

It also reveals that the naturalness of the unit selection TTS was just slightly superior to that of the HMM-based TTS when evaluating with very long sentences due to its unstability on segmental duration and timing. The naturalness of the HTT also slightly outperformed the unit selection and HMM-based TTS. Therefore, although the naturalness of HTT is significantly high with huge amount of data for rendering [4], it is significantly reduced under limited data conditions.

The results from the intelligibility evaluation were consistent with the results from a HMM-based TTS [13] 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 score of speech analyzed / synthesized by MRTD-STRAIGHT was lower than that of the original speech due to the reconstruction errors of MRTD and STRAIGHT. The MOS scores for speech synthesized by all synthesizers were not compared with that of speech analyzed / synthesized by MRTD-STRAIGHT and the original speech, which are the two ideal limitations of HTD and unit selection TTS, HTT since they were implemented under a “limited data condition”.

3.4. Discussions on differences between the proposed HTD and HTT

Although the proposed HTD shares some common procedures with HTT [4], their concepts are completely different. These

Table 1: Word Error Rates (%)

	HMM	HTT	Unit Selection	HTD	MRTD-STRAIGHT	Original
Mean	0.25	3.82	10.83	0.25	0.25	0
95% confidence	0.09	0.71	0.88	0.09	0.09	0

differences are presented and discussed in this section. There are three main differences:

(1) HTT replaces all frames of the guided trajectory generated by HMMs with the closest frames found in the original database. Therefore, HTT can be considered to be one kind of unit selection that uses HMM-based TTS as an intermediate procedure 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, their high computational load, and their inflexibility for voice transformations.

The proposed HTD uses HMM-based TTS to generate spectral and prosodic trajectories. The spectral trajectory is then decomposed into its event functions and event targets. The prosodic trajectories and the event functions of the spectral trajectory are preserved to maintain the high intelligibility of HMM-based TTS, while the sparse event targets are replaced with the event targets located at the closest frames found in the original database to reduce over-smoothness in spectral sequence. Therefore, the proposed HTD is one extended version of HMM-based TTS in which speech synthesized by HMMs is transformed to the original speech by using MRTD, resulting in the improvement of synthesized speech in terms of naturalness while preserving main advantages of HMM-based TTS.

(2) HTT requires a huge database for rendering to ensure the smoothness of the synthesized speech because limited data may cause mismatches and discontinuities between consecutive frames. The smoothness of the synthesized trajectory in the proposed 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, the proposed HTD can synthesize stable and smooth speech even under limited data conditions.

(3) HTT can be combined with voice transformation by using multiple huge target databases for rendering. 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 [7], [8] 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 spectral sequence synthesized by HMM-based TTS are preserved and its event targets are selected from an original database, or are transformed to those of the original speech. Therefore, it is possible to develop the proposed HTD to synthesize multiple voices with a multiple-voices database. The experimental results in this paper revealed that the proposed HTD was efficient with a small database. Therefore, the proposed HTD can be developed for voice transformations with limited target data. Although the proposed HTD was just evaluated with a single-voice database, it will be implemented with multiple-speakers and multiple-styles databases in the future to confirm its flexibility for voice transformations.

4. Conclusions

In this paper, a hybrid TTS among unit selection, HMM-based TTS, and MRTD, named HTD, was proposed. The experimental results show that the proposed HTD could borrow both the high intelligibility of HMM-based TTS and the high naturalness of unit selection TTS under limited data conditions. In the future, we will investigate other possible advantages of the proposed HTD such as the flexibility for voice transformations. We will also implement the proposed method with other languages to confirm the unification and language-independence of the proposed TTS.

5. 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).

6. References

[1] A.J. Hunt, A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP 96,* **1**, pp. 373–376, (1996).

[2] H. Zen, T. Toda "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," *Proc. Interspeech,* (2005).

[3] T. Toda, K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. and Syst.,* Vol. E90-D, Issue 5, pp. 816-824, (2007).

[4] Y. Qian, F. K. Soong, Z. Yan, "A Unified Trajectory Tiling Approach to High Quality Speech Rendering," *IEEE Trans. on Audio, Speech, and Language Proc.,* Vol. 21, No. 2, pp. 280–290, (2013).

[5] R.B. Chicote, J. Yamagishi, S. King, J.M. Montero, J.M. Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech communication,* Vol. 52, pp. 394404, (2010).

[6] PC. Nguyen, T. Ochi and M. Akagi, "Modified restricted temporal decomposition and its application to low rate speech coding," *IEICE Trans. Inf. and Syst.,* E86-D3 (2003).

[7] P.N. Binh and M. Akagi, "Efficient modeling of temporal structure of speech for applications in voice transformation," *Interspeech 2009,* pp. 1631-1634, (2009).

[8] V. Popa, J. Nurminen, M. Gabbouj, "A Novel Technique for Voice Conversion Based on Style and Content Decomposition with Bilinear Models," *Interspeech 2009*, pp. 2655-2658, (2009).

[9] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP-83,* pp. 81–84 (1983).

[10] H. Kawahara, "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci & Tech.,* **27**(6), 349–353 (2006).

[11] H. Phe, *Chinh ta Tieng Viet (Vietnamese Grammar),* (Da Nang Publisher), pp. 9–15, (2003).

[12] L.C. Mai and D.N. Duc, "Design of Vietnamese speech corpus and current status," *Proc. ISCSLP-06*, pp. 748–758 (2006).

[13] TT. Vu, MC. Luong and S. Nakamura, "An HMM-based Vietnamese speech synthesis system, Speech Database and Assessments," *Proc. COCOSDA-2009,* pp. 116–121 (2009).

[14] T.V. Do, D.D. Tran, and T.T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," *Proc. SoICT '11,* pp. 165-171, (2011).