A SPEAKING RATE-CONTROLLED MANDARIN TTS SYSTEM

Chiao-Hua Hsieh[1], Yih-Ru Wang[1], Chen-Yu Chiang[2], Sin-Horng Chen[1]

[1] Department of Electrical Engineering, National Chiao Tung University,
[2] Department of Communication Engineering, National Taipei University,

ABSTRACT

In this paper, a new speaking rate-controlled Mandarin TTS system based on a speaking rate-dependent hierarchical prosodic model (SR-HPM) [6] is proposed. In the training phase, a data-driven approach is employed to automatically build the SR-HPM directly from a large prosody-unlabeled speech database containing utterances of various speaking rates. The SR-HPM comprises 15 sub-models designed to describe various relationships among 3 types of prosodic-acoustic features of speech utterances, two types of prosodic tags specifying a 4-layer prosody hierarchy, linguistic features of various levels of the associated texts, and the speaking rates. In the test phase, the SR-HPM is employed to generate 4 prosodic-acoustic features, including syllable pitch contours, syllable durations, syllable energy levels, and syllable juncture pause durations. Combining these prosodic features with the spectral features generated by the HTS synthesizer, the system can generate natural speech for any speaking rate in a wide range of 0.15-0.3 seconds/syllable. A distinct feature of the system to control the occurrence frequencies of breaks of various types as well as their pause durations according to the given speaking rate was demonstrated. A subjective test showed that MOS scores of 3.35, 3.44 and 3.28 were achieved respectively for fast (SR=0.17 sec/syllable), medium (SR=0.2 sec/syllable) and slow (SR=0.25 sec/syllable) synthetic speeches.

Index Terms— Speaking rate modeling, Mandarin prosody modeling, Speaking rate-controlled TTS

1. INTRODUCTION

Speaking rate (SR) control for TTS is to adjust the speed of the synthesized speech. A good speaking rate control scheme can make the synthesized speech sound more vividly to away from the criticism of machine-like sounding [1-3] as well as be suitable for some special applications, e.g. fast rate for people with vision disability [4-5] and slow rate for elder people.

Conventionally, speaking rate control is realized via adjusting the prosodic features of the synthesized speech. Most existing TTS systems implement the function by adjusting the durations of synthesis units, e.g. syllables or words, as well as pause durations. In [1], the effects of speaking rate adjustment on prosodic features in Chinese TTS were investigated. The study focused on prosodic structure variation, duration variation, F0 distribution and variation, and accent placement variation. A phoneme duration control method for HMM-based Japanese TTS by interpolating models of fast, normal and slow speech rates was proposed in [2]. The subjective performance of three speech rate control methods for HMM-based speech synthesis is evaluated in [3]. It found that the scheme using proportional duration adjustment performed well for slow speech synthesis. In [4], the intelligibility of ultra-fast TTS synthesized speech for screen reading software used by persons with visual disability was studied. A method to generate synthetic speech at high speaking rates based on interpolating hidden semi-Markov models trained on speech corpora of normal and fast speaking rates was proposed in [5]. The synthesized speech was evaluated by blind and sighted listeners.

In this paper, a speaking rate-controlled Mandarin TTS system based on a speaking rate-dependent hierarchical prosodic model (SR-HPM) [6] is proposed. Via designing and training a sophisticated SR-HPM describing the influences of speaking rate on Mandarin speech prosody in the training phrase, we can use the SR-HPM in the test phase to generate all the prosodic features required to synthesize a nature output speech for a given input text and a designated speaking rate.

The remaining part of the paper is organized as follows. Section 2 presents the proposed speaking rate-controlled Mandarin TTS system in detail. Experimental results of realizing the system using a speech corpus of a female speaker is discussed in Section 3. Some conclusions are given in the last section.

2. THE PROPOSED SYSTEM

Fig.1 displays a block diagram of the proposed system. It consists of two phases: training and test. In the training phase, an SR-HPM is trained from a large speech corpus comprising prosody-unlabeled utterances of various speaking rates. Here, speaking rate is defined as the average syllable duration of an utterance without considering inter-syllable pauses. In the test phase, the SR-HPM is employed to firstly predict the break types of all syllable junctures, and to then predict the three prosodic states of each syllable. All prosodic-acoustic features required to synthesize the output speech are then generated by the SR-HPM and SR-specific feature denormalization functions. Lastly, synthetic speech is generated by combining the predicted prosodic-acoustic features and the spectral features generated by an HMM-based speech synthesizer [7]. Some parts of the system are discussed in more detail in the following subsections.

2.1 SR-HPM Training

In the training phase, an existing method proposed previously [6] is employed to train the SR-HPM from a prosody-unlabeled speech corpus containing utterances of various speaking rate. Fig.2 shows a block diagram of the method. It first constructs some speaking rate-specific normalization functions for the four prosodic-acoustic features concerned in this study, viz syllable duration, syllable pitch contour, syllable energy level, and syllable juncture pause duration, from all utterances of the whole database.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Here, the first three features are Z-normalized, while the last one is *cdf*-normalized based on Gamma distribution. Utterance-based normalizations for these four prosodic-acoustic features are applied using these functions. A modified PLM (prosody labeling and modeling) algorithm [6,8] is then employed to construct the SR-HPM containing 15 sub-models designed to describe various relationships among speaking rate and 3 types of features, viz 4 prosodic-acoustic features of speech utterances, two types of prosodic tags representing a 4-layer prosody hierarchy [8,9], and various linguistic features of the associated texts. The modified PLM algorithm is a sequential optimization procedure. With proper initialization to firstly generate initial parameters of all sub-models and label all prosodic tags, it updates model parameters and re-labels prosodic tags sequentially and iteratively until a convergence is reached.

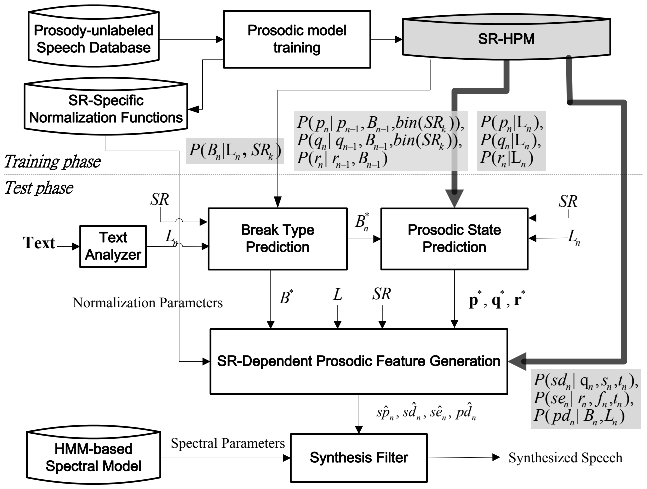

Prosody-unlabeled Speech Database
Prosodic model training
SR-HPM
SR-Specific Normalization Functions
$P(B_n|L_n, SR_k)$
$P(p_n| p_{n-1}, B_{n-1}, bin(SR_k))$, $P(q_n| q_{n-1}, B_{n-1}, bin(SR_k))$, $P(r_n| r_{n-1}, B_{n-1})$
$P(p_n|L_n)$, $P(q_n|L_n)$, $P(r_n|L_n)$
Training phase
Test phase
SR
Text
Text Analyzer
L_n
Break Type Prediction
B_n^*
Prosodic State Prediction
SR
L_n
Normalization Parameters
B^*
L
SR
$\mathbf{p}^*, \mathbf{q}^*, \mathbf{r}^*$
SR-Dependent Prosodic Feature Generation
$P(sd_n| q_n, s_n, t_n)$, $P(se_n| r_n, f_n, t_n)$, $P(pd_n| B_n, L_n)$
$\hat{sp}_n, \hat{sd}_n, \hat{se}_n, \hat{pd}_n$
HMM-based Spectral Model
Spectral Parameters
Synthesis Filter
Synthesized Speech

Fig. 1: A block diagram of the proposed speaking rate-controlled Mandarin TTS system.

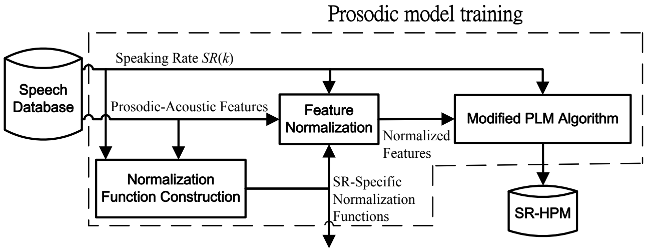


Fig. 2. A schematic diagram of the proposed speaking rate modeling approach.

Among these 15 sub-models, 11 are used in the test phase. They are described in more detail as follows. Firstly, four prosodic-acoustic sub-models used to describe the variations of syllable pitch contour, syllable duration, syllable energy level, and syllable juncture pause duration are expressed by

$$P(\mathbf{sp}_n \mid p_n, B_{n-1}^n, t_{n-1}^{n+1}) = N(\mathbf{sp}_n; \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n} + \boldsymbol{\beta}^f_{B_{n-1}, tp_{n-1}} + \boldsymbol{\beta}^b_{B_n, tp_n} + \boldsymbol{\mu}_{sp}, R_{sp}) \tag{1}$$

$$P(sd_n \mid q_n, s_n, t_n) = N(sd_n; \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + \mu_{sd}, R_{sd}) \tag{2}$$

$$P(se_n \mid r_n, f_n, t_n) = N(se_n; \omega_{t_n} + \omega_{f_n} + \omega_{r_n} + \mu_{se}, R_{se}) \tag{3}$$

$$P(pd_n|B_n, L_n) = g(pd_n; \alpha_{B_n, L_n}, \beta_{B_n, L_n}) \tag{4}$$

where $\mathbf{sp}_n$ is a 4-dimensional pitch contour feature vector [10]; $\boldsymbol{\beta}$'s are affecting patterns (APs) to count the respective influences of the affecting factors (AFs) of tone t_n, prosodic state p_n, tone pairs $tp_{n-1} = (t_{n-1}, t_n)$ and tp_n, breaks B_{n-1} and B_n on the pitch contour variation; $\boldsymbol{\mu}_{sp}$ and R_{sp} are global mean vector and residual's variance; sd_n is syllable duration, and γ 's are its APs with AFs of tone t_n, base-syllable s_n, and prosodic state q_n ,; se_n is syllable energy level, and ω 's are its APs with AFs of t_n, final type f_n, and prosodic state r_n; pd_n is pause duration, g is a Gamma *pdf* with two parameters α_{B_n, L_n} and β_{B_n, L_n} depending on break type B_n and context L_n. Here $g(pd_n; \alpha_{B_n, L_n}, \beta_{B_n, L_n})$ is realized as a decision tree formed by the CART algorithm [11].

Secondly, the break-syntax sub-model, $P(B_n \mid L_n, SR_k)$, is used to describe the dependence of the occurrence of B_n on L_n and speaking rate SR_k. It is also realized as a decision tree formed by the CART algorithm [11]. Thirdly, three prosodic state sub-models of $P(p_n \mid p_{n-1}, B_{n-1}, bin(SR_k))$, $P(q_n \mid q_{n-1}, B_{n-1}, bin(SR_k))$ and $P(r_n \mid r_{n-1}, B_{n-1})$ are used to describe the transition probabilities of the three prosodic states, including p_n for syllable pitch contour, q_n for syllable duration, and r_n for syllable energy level. Here, $bin(SR_k)$ denote a bin of the histogram of SR_k. Lastly, three prosodic state-syntax sub-models of $P(p_n \mid L_n)$, $P(q_n \mid L_n)$ and $P(r_n \mid L_n)$ are used to describe the dependencies of the occurrences of p_n, q_n, and r_n on L_n. They are also realized by the CART algorithm [11].

2.2 The Test Phase

The test phase is designed to generate the four prosodic-acoustic features of syllable pitch contour, syllable duration, syllable energy level and syllable-juncture pause duration using given speaking rate *SR* and linguistic features extracted from the input text. As shown in Fig. 1, the break type of each syllable juncture is predicted firstly by

$$B_n^* = \arg\max_{B_n} P(B_n \mid L_n, SR) \tag{5}$$

Then, the three prosodic states of each syllable are predicted by

$$\begin{aligned} &\mathbf{p}^*,\ \mathbf{q}^*,\ \mathbf{r}^* \\ &= \arg\max_{\mathbf{p},\mathbf{q},\mathbf{r}} \big(P(p_1|\text{bin}(SR))P(q_1|\text{bin}(SR))P(r_1) \\ &\quad \cdot \prod_{n=2}^{N} \begin{pmatrix} P(p_n \mid p_{n-1}, B_{n-1}^*, \text{bin}(SR)) \\ \cdot P(q_n \mid q_{n-1}, B_{n-1}^*, \text{bin}(SR)) \\ \cdot P(r_n \mid r_{n-1}, B_{n-1}^*) \end{pmatrix} \\ &\quad \cdot \left(\prod_{n=1}^{N} P(p_n \mid L_n) P(q_n \mid L_n) P(r_n \mid L_n) \right) \end{aligned} \tag{6}$$

The normalized versions of the four prosodic-acoustic features are then generated using the predicted break types and prosodic states by

$$\mathbf{sp}_n^* = \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n^*} + \boldsymbol{\beta}^f_{B_{n-1}^*, t_{n-1}^n} + \boldsymbol{\beta}^b_{B_n^*, t_n^{n+1}} + \boldsymbol{\mu}_{sp} \tag{7}$$

$$sd_n^* = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n^*} + \mu_{sd} \tag{8}$$

$$se_n^* = \omega_{t_n} + \omega_{f_n} + \omega_{r_n^*} + \mu_{se} \tag{9}$$

$$pd_n^* \equiv \mu_n^* = \alpha_n^* \beta_n^* \tag{10}$$

Here $\mathbf{sp}_n^*$, sd_n^* and se_n^* are generated using the syllable-based prosodic-acoustic sub-models without the residual terms. Since most residuals are quite small, their neglects do not cause big trouble. The pause duration pd_n^* is generated from the mean of the Gamma distribution with parameters α_n^* and β_n^* found from the leaf node of the break-acoustic sub-model specified by the predicted break type B_n^* and the contextual features L_n.

Lastly, the final four prosodic-acoustic features are obtained by performing the denormalization operations to $\mathbf{sp}_n^*$, sd_n^*, se_n^* and pd_n^* using the inverse functions of the normalization functions found in the training of the SR-HPM. Specifically,

$$\hat{\mathbf{sp}}_n(i) = \frac{\mathbf{sp}_n^*(i) - \mu_g^{sp}(t_n,i)}{\sigma_g^{sp}(t_n,i)}\tilde{\sigma}^{sp}(SR,t_n,i) + \tilde{\mu}^{sp}(SR,t_n,i) \quad \text{for } i = 1,2,3,4 \tag{11}$$

$$s\hat{d}_n = \frac{sd_n^* - \mu_g^{sd}}{\sigma_g^{sd}}\tilde{\sigma}^{sd}(SR) + \mu_k^{sd} \tag{12}$$

$$s\hat{e}_n = se_n^* \tag{13}$$

$$p\hat{d}_n = \mathrm{G}^{-1}(\mathrm{G}(pd_n^*, \alpha_g^{pd}, \beta_g^{pd}), \tilde{\alpha}^{pd}(SR), \tilde{\beta}^{pd}(SR)) \tag{14}$$

Here, G denotes the cdf of Gamma distribution.

It is noted that $\mu_k^{sd} = SR$ and the reconstructed energy level is not SR-dependent. Using these prosodic feature estimates, we can combine them with the spectral parameters generated from the HMM-based speech synthesizer [7] to generate the synthetic speech.

3. EXPERIMENTAL RESULTS

The speaking rate-controlled Mandarin TTS system was implemented using a prosody-unlabeled speech database. The database is composed of four parallel speech corpora of a female professional announcer with fast, normal, medium and slow speaking rates. Fig. 3 shows the histogram (utterance count) of speaking rate in seconds/syllable of the four corpora. All utterances are short paragraphs. There are in total 1,478 utterances comprising 203,746 syllables. As shown in the figure, the SRs of utterances in these four speech corpora distribute widely in the range of 0.147-0.297 seconds/syllable (or 3.4-6.8 syllables/second) and overlap seriously. We divided the database into two sets, one for training and one for testing. These two sets contained 183,795 and 19,951 syllables, respectively. Besides, the speech corpus with normal speaking rate was used to train an HMM-based speech synthesizer [7] for spectral feature generation.

The training set was used to build the SR-HPM. Then, the well-trained SR-HPM was firstly used in the test phase to predict the break types of all syllable junctures by Eq.5. Table I(a) shows the confusion matrix of the break type prediction for the test set. It can be seen from the table that the predictions for *B*0, *B*1 and *B*4 were good, while all others were fair or poor. The overall accuracy was 71.1%. If we used the conventional four-level break tags with *BT*1={*B*0,*B*1, *B*2-1, *B*2-3}, *BT*2={*B*2-2}, *BT*3={*B*3} and *BT*4={*B*4}, the performance was 86.1%. Its confusion table is showed in Table I(b). This performance is comparable to that of the method proposed in [12] which was evaluated on a simpler task of three-level break prediction using sentential utterances of normal speaking rate. Precision and recall rates of 93.17 and 93.07 for non-breaks, 76.24 and 79.82 for PW breaks, and 67.78 and 61.41 for PP breaks were achieved in [12].

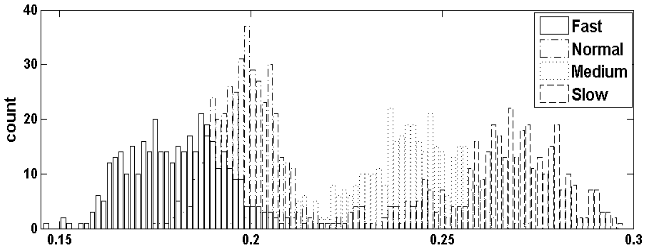


Fig. 3. Histogram of utterance's speaking rate of the database formed by four parallel corpora.

TABLE I
THE CONFUSION MATRIX OF THE BREAK TYPE PREDICTION FOR THE TEST SET: (A) 7-LEVEL BREAK TYPES AND (B) 4-LEVEL BREAK TYPES.

(a)

Tar\Pre	*B*0	*B*1	B2-1	B2-2	*B*2-3	*B*3	*B4*	Total
*B*0	**86.8**	8.9	2.4	1.6	0.2	0.0	0.0	3034
*B*1	4.6	**87.0**	4.9	2.9	0.5	0.2	0.0	9506
*B*2-1	7.1	34.2	**40.9**	15.7	1.5	0.9	0.0	2258
*B*2-2	5.4	9.2	15.8	**55.6**	0.5	13.3	0.2	1985
*B*2-3	7.3	39.2	22.3	25.3	**4.1**	1.9	0.0	1076
*B*3	1.8	2.2	3.7	17.5	0.0	**38.8**	36.0	1218
*B*4	0.0	0.0	0.2	0.6	0.0	13.4	**85.8**	754
Average =71.1								

(b)

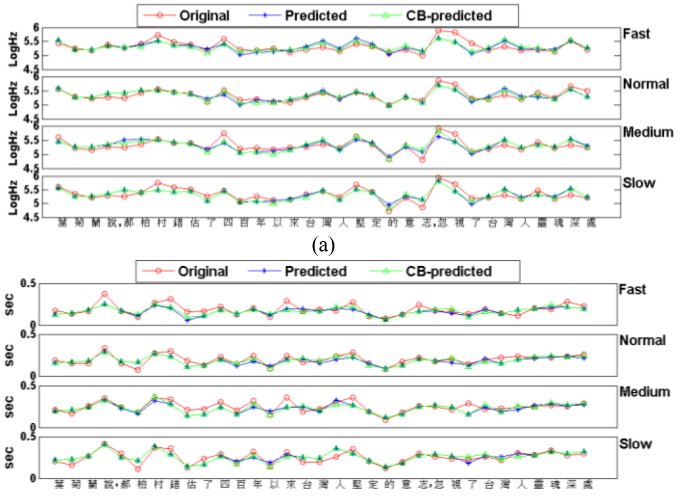
Tar\Pre	*BT1*	*BT2*	*BT3*	*BT4*	Total
BT1	**93.6**	6.0	0.4	0.0	15874
BT2	30.9	**55.6**	13.3	0.2	1985
BT3	7.7	17.5	**38.8**	36.0	1218
BT4	0.2	0.6	13.4	**85.8**	754
Ave =86.1					

We then predicted the three types of prosodic states for all syllables by Eq.6. The four prosodic-acoustic features were then generated by Eqs.7-10 using the predicted break types and prosodic states as well as some linguistic features. They were then denormalized by Eqs.11-14 to generate the final values. Table II displays the root-mean-squared errors (RMSEs) of the four prosodic-acoustic feature estimates for the test set. RMSEs of 48.9 ms, 0.19 log-Hz, 3.64 dB, and 88.5 ms were achieved respectively for syllable duration, syllable pitch contour, syllable energy level, and syllable-juncture pause duration. We also find from the table that, except the pause duration, these values were insensitive to the break type prediction error. This mainly resulted from the fact that only few serious break type prediction errors, say between non-pause breaks of (*B*0, *B*1, *B*2-1, *B*2-3) and long-pause breaks of (*B*3, *B*4), were found from Table I. By a more detailed analysis, we found that most large pause duration errors resulted from the pair-wise confusions of (*B*3, *B*4) and (*B*2-2, *B*3) instead of (*BT*1, *B*3/*B*4). Actually, these two types of confusions were not perceptually annoying. So, these objective performances were reasonably good.

TABLE II
RMSEs OF THE FOUR PROSODIC-ACOUSTIC FEATURE ESTIMATES.

	Using predicted breaks and prosodic states	Using correct breaks and predicted prosodic states
Syllable duration	48.9 ms	48.2 ms
Syllable pitch contour	0.18 log-Hz	0.17 log-Hz
Syllable energy level	3.64 dB	3.55 dB
Juncture pause duration	88.5 ms	55.0 ms

A typical example of the estimated syllable pitch level and syllable duration for four speaking rates of fast (*SR*=0.19), normal (*SR*=0.20 sec/syllable), medium (*SR*=0.24 sec/syllable), and slow (*SR*=0.27 sec/syllable) is shown in Fig. 4. It can be found from these two figures that most estimated values matched well with their original counterparts. So, the predictions of these two features were reasonably good.



(a)

(b)

Fig. 4. An example of estimated (a) syllable pitch level and (b) syllable duration for four speaking rates of fast (*SR*=0.18), normal (*SR*=0.20), medium (*SR*=0.24) and slow (*SR*=0.26).

An example of speech synthesis for a paragraph with 8 different speaking rates is given.[1] Fig. 5 displays the break type predictions and their pause duration estimates for parts of these 8 synthesized utterances. It can be found from the figure that not only more short- to long-pause breaks (i.e., *B*2-2, *B*3, *B*4) were found as *SR* increased, but also their pause durations increased as *SR* increased. This result matched well with our prior knowledge about the relationship between syllable juncture break pause and speaking rate. An informal listening test confirmed that all these 8 synthesized utterances sounded very nature. To the best of our knowledge, this sophisticated pause generation is a distinct feature of the proposed system not founded in all other existing systems.

Lastly, a subjective test was performed to examine the naturalness of the synthesized speech. For comparison, the speaking rate control scheme [13] proposed for the HTS system was also tested and taken as baseline. Fifteen subjects were involved in the test. They were all graduate students. Ten short paragraphs with length from 35 to 70 syllables were selected from the outside test data set. Three speaking rates of fast (SR=0.17 sec/syllable), medium (SR=0.20 sec/syllable), and slow (SR=0.25 sec/syllable) were tested. In the test, each subject was asked to give the MOS score to each of these synthesized utterances by the proposed system and by the HTS scheme prompted to the listener in random order. The original speech was always provided to the subject for his reference. Table III lists the resulting MOS scores. We find from the table that the MOS scores of the two systems were comparable for the medium speaking rate. For both fast and slow speaking rates, the MOS scores of the proposed system degraded slightly while they degraded seriously for the HTS scheme. So, the proposed system significantly outperformed the HTS scheme on the function of speaking rate control. We can therefore conclude that the proposed speaking rate-control method performed very well.

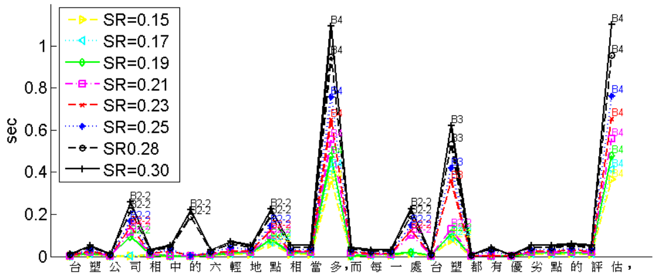


Fig. 5. An example of the break type predictions and their pause duration generations for parts of 8 synthesized utterances.

TABLE III
THE RESULTS OF SUBJECTIVE TEST.

MOS	Fast (*SR*=0.17)	Medium (*SR*=0.20)	Slow (*SR*=0.25)
HTS	2.44	3.36	2.7
proposed system	3.35	3.44	3.28

4. CONCLUSIONS

A speaking rate-controlled Mandarin TTS system designed based on the SR-HPM has been realized. The proposed system has showed to have good prosody generation capability. A distinct feature of the system to control the occurrence frequencies of different break types as well as their pause durations according to the given speaking rate was demonstrated. High performance of the speaking rate control function of the system was confirmed by a subjective test.

ACKNOWLEDGMENTS

This work was supported in part by the NSC of Taiwan under Contract NSC99-2221-E-009-009-MY3, and in part by the MoE ATU plan. The authors would like to thank the ACLCLP for providing the Treebank Corpus.

[1] Y. Zu, A. Li and Y. Li, "Speech Rate Effects on Prosodic Features," Report of Phonetic Research 2006, Institute of Linguistics, Chinese Academy of Social Sciences, pp. 141-144.

[2] K. Iwano, M. Yamada, T. Togawa, and S. Furui, "Speech-rate variable HMM-based Japanese TTS system," in Proc. TTS2002, Sept. 2002.

[1] The examples was available at https://dl.dropbox.com/u/28129441/example_waveform/mm.tar.gz.

[3] T. Kato, M. Yamada, N. Nishizawa, K. Oura and K. Tokuda, “Large-scale Subjective Evaluations of Speech Rate Control Methods for HMM-based Speech Synthesizers,” in Proc. INTERSPEECH-2011, Aug. 2011, pp. 1845-1848.

[4] T. Nishimoto, S. Sako, S. Sagayama, K. Ohshima, K. Oda and T. Watanabe, “Effect of Learning on Listening to Ultra-Fast Synthesized Speech,” in Proc. EMBC2006, Sept. 2006, pp. 5691-5694.

[5] M. Pucher, D. Schabus and J. Yamagishi, “Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners,” in Proc. INTERSPEECH-2010, Sept. 2010, pp. 2186-2189.

[6] C. H. Hsieh, C. Y. Chiang, Y. R. Wang, H. M Yu, S. H. Chen, “A New Approach of Speaking Rate Modeling for Mandarin Speech Prosody,” in Proc. INTERSPEECH-2012, Sept. 2012

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, in Proc. ICASSP’00, Jun. 2000, pp.1315-1318.

[8] C. Y. Chiang, S. H. Chen, H. M. Yu, and Y. R. Wang, “Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech,” J. Acoust. Soc. Am., vol. 125, No. 2, pp. 1164-1183, Feb, 2009.

[9] C. Y. Tseng, S. H. Pin, Y. L. Lee, H. M. Wang, and Y. C. Chen, “Fluent speech prosody: Framework and modeling,” Speech Commun., vol. 46, no. 3-4, pp. 284-309, 2005.

[10] S. H. Chen and Y. R. Wang, “Vector quantization of pitch information in Mandarin speech,” IEEE Trans. Commun., vol. 38, no. 9, pp. 1317-1320, Sept. 1990.

[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Tree. Wadsworth, Belmont, 1984.

[12] Chi-Chun Hsia, Chung-Hsien Wu, and Jung-Yun Wu, “Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis,” IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 8, pp.1994-2003, August 2010.

[13] T. Yoshimura, Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems, Ph.D thesis, Nagoya Institute of Technology, Jan. 2002.