# Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS

Sin-Horng Chen, *Senior Member, IEEE*, Chiao-Hua Hsieh, Chen-Yu Chiang, *Member, IEEE*, Hsi-Chun Hsiao, Yih-Ru Wang, *Member, IEEE*, Yuan-Fu Liao, and Hsiu-Min Yu

***Abstract*—A new data-driven approach to building a speaking rate-dependent hierarchical prosodic model (SR-HPM), directly from a large prosody-unlabeled speech database containing utterances of various speaking rates, to describe the influences of speaking rate on Mandarin speech prosody is proposed. It is an extended version of the existing HPM model which contains 12 sub-models to describe various relationships of prosodic-acoustic features of speech signal, linguistic features of the associated text, and prosodic tags representing the prosodic structure of speech. Two main modifications are suggested. One is designing proper normalization functions from the statistics of the whole database to compensate the influences of speaking rate on all prosodic-acoustic features. Another is modifying the HPM training to let its parameters be speaking-rate dependent. Experimental results on a large Mandarin read speech corpus showed that the parameters of the SR-HPM together with these feature normalization functions interpreted the effects of speaking rate on Mandarin speech prosody very well. An application of the SR-HPM to design and implement a speaking rate-controlled Mandarin TTS system is demonstrated. The system can generate natural synthetic speech for any given speaking rate in a wide range of 3.4-6.8 syllables/sec. Two subjective tests, MOS and preference test, were conducted to compare the proposed system with the popular HTS system. The MOS scores of the proposed system were in the range of 3.58-3.83 for eight different speaking rates, while they were in 3.09-3.43 for HTS. Besides, the proposed system had higher preference scores (49.8%-79.6%) than those (9.8%-30.7%) of HTS. This confirmed the effectiveness of the speaking rate control method of the proposed TTS system.**

***Index Terms*—Mandarin prosody modeling, speaking rate modeling, speaking rate-controlled TTS.**

Manuscript received December 10, 2013; revised February 27, 2014; accepted April 15, 2014. Date of publication May 02, 2014; date of current version May 16, 2014. 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 associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

S. H. Chen, C. H. Hsieh, H. C. Hsiao, and Y. R. Wang are with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: schen@mail.nctu.edu.tw; yrwang@mail.nctu.edu.tw).

C. Y. Chiang is with the Department of Communication Engineering, National Taipei University, Taipei 23741, Taiwan.

Y. F. Liao is with the Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan.

H.-M. Yu is with the Language Center, Chung Hua University, Hsinchu 30012, Taiwan.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes synthesized speech examples of both the proposed system and the baseline system with eight different speaking rates. This material is 17.2 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2321482

## I. Introduction

SPEAKING RATE (SR) is a prosodic feature that influences many speech phenomena such as syllable duration, pause duration, prosodic phrasing, occurrence frequency of pause, word pronunciation, phone contraction, pitch contour shape, and so on. Exploring the effects of speaking rate on prosodic/linguistic features [1], [2] are interesting research issues. [1] investigated the effects of speech rate on discourse prosody of Chinese speech and concluded that the effects are nonlinear. [2] explored the effect of speech rate on prosodic phrasing in Korean speech; and found that accentuated phrase includes 5 or fewer syllables at normal rate, but can include up to 7 syllables at fast rate.

Modeling the effects of speaking rate is also an important research issue in both automatic speech recognition (ASR) and text-to-speech (TTS). For ASR, the main concern is how to compensate the speaking rate effect in order to improve the relatively-low recognition performance of fast or slow speech [3]–[9]. Methods proposed included speaking rate normalization of spectral feature [3], [4], use of durational information [5], modeling of pronunciation variation [6], adjustment of mixture weights and transition probabilities [7], use of parallel rate-specific acoustic models [8], and decoding strategy adaptation [9]. For TTS, the speaking rate control of the synthetic speech is needed for making it sound more vividly to away from the criticism of machine-like sounding [10]–[17] as well as for being suitable for some special applications, e.g. fast rate for people with vision disability [18], [19]. Methods proposed included proportional duration adjustment [11], modeling of speech rate effects on prosodic features [12]–[14], model interpolation [10], [15], [19], and phone/syllable duration modeling [16], [17]. Besides, speaking rate change was also considered in voice conversion [20].

We find from those previous studies that an unsolved issue is the lack of a systematic way to build a quantitative model to account for all major influences of speaking rate on speech prosody so as to be used in various applications. In this study, we adopt a new approach to solve the problem based on an existing prosody labeling and modeling (PLM) algorithm which builds a sophisticated hierarchical prosodic model (HPM) of Mandarin speech containing 12 sub-models to describe various relationships of the prosodic-acoustic features of speech signal, the linguistic features of the associated text, and the prosodic tags representing a 4-layer prosody structure of the utterance [21]. The current speaking rate modeling approach takes two strategies to modify the PLM algorithm in order to automatically generate a speaking rate-dependent HPM (SR-HPM) from a

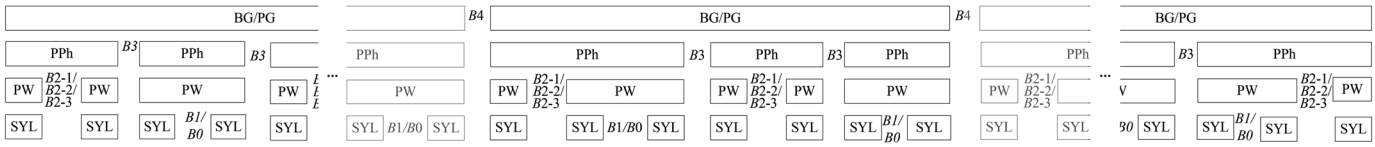



Fig. 1. The hierarchical structure of Mandarin prosody used in this study [21].

large prosody-unlabeled speech database containing utterances of various speaking rates. One is taking the speaking rate as a continuous independent variable to design proper feature normalization functions from the statistics of the whole database in order to normalize the prosodic-acoustic features for compensating the influences of speaking rate in 7 prosodic- acoustic feature-related sub-models of the SR-HPM. Another is modifying the training procedure of the PLM algorithm to let the parameters of 3 other sub-models be speaking rate dependent. The approach is in contrast to our previous study of realizing a speaking rate-controlled TTS system via model interpolation using four HPM models trained from four parallel speech corpora of a female speaker with fast, normal, medium and slow speaking rates [10]. The current study builds a single sophisticated SR-HPM to more accurately describe the influences of speaking rate on Mandarin speech prosody from an aggregation of the same four parallel speech corpora. Using the SR-HPM, a better speaking rate-controlled TTS system can be built.

Several advantages of the proposed approach can be found. First, the influences of speaking rate on Mandarin speech prosody can be automatically learned from a large database without human's prosody labeling. Second, the effects of speaking rate on many important prosodic phenomena, such as prosodic-acoustic feature variations, prosodic phrasing and occurrence frequencies of breaks, can be directly investigated from the SR-HPM model. Third, since speaking rate becomes a continuous independent variable of the SR-HPM model, it is easy to consider the effect of speaking rate in some applications, such as ASR and TTS, via using the SR-HPM. In this study, a speaking rate-controlled Mandarin TTS system is realized to demonstrate such an application.

The paper is organized as follows. Section II gives a brief review of the existing HPM model and the PLM algorithm proposed previously. Section III presents the proposed speaking rate modeling approach to generate the SR-HPM model in detail. Experimental results of the speaking rate modeling are also discussed. An application of the SR-HPM to design and implement a speaking rate-controlled Mandarin TTS system is demonstrated in Section IV. Some conclusions are given in the last section.

## II. Review of the Existing HPM Model

The HPM [21] is a statistical prosodic model designed to describe various relationships of prosodic-acoustic features $\mathbf{A}$, tags of prosody structure $\mathbf{T}$, and linguistic features $\mathbf{L}$. Three types of prosodic-acoustic features are modeled, including syllable-based features $\mathbf{X}$, syllable juncture-based features $\mathbf{Y}$, and inter-syllable differential prosodic-acoustic features $\mathbf{Z}$. Here, $\mathbf{X}$ includes syllable pitch contour feature vector $\mathbf{sp}_n = [\, a_n^0 \quad a_n^1 \quad a_n^2 \quad a_n^3 \,]^T$ which contains four coefficients of a 3-rd order orthogonal polynomial expansion [22], syllable duration $sd_n$, and syllable energy level $se_n$ of the $n$-th syllable; $\mathbf{Y}$ includes pause duration $pd_n$ and energy-dip level of the syllable juncture between the $n$-th and $(n+1)$-th syllables (referred to as juncture $n$); and $\mathbf{Z}$ includes a normalized pitch-level jump $pj_n$ and two normalized duration lengthening factors, $dl_n$ and $df_n$, across juncture $n$. So, the complete prosodic-acoustic feature sequence is $\mathbf{A} = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$; where $\mathbf{X} = \{\mathbf{sp}, \mathbf{sd}, \mathbf{se}\}$, $\mathbf{Y} = \{\mathbf{pd}, \mathbf{ed}\}$ and $\mathbf{Z} = \{\mathbf{pj}, \mathbf{dl}, \mathbf{df}\}$ represent sequences of the above prosodic-acoustic features.

The prosody structure considered in the HPM is a four-layer prosody hierarchy shown in Fig. 1. It is a modified version of the hierarchical prosodic phrase grouping (HPG) model proposed by Tseng [23]. It is composed of four types of layered prosodic constituents: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath/prosodic phrase group (BG/PG). The prosody hierarchy is represented in terms of two types of prosody tags $\mathbf{T} = \{\mathbf{B}, \mathbf{P}\}$: the break type $\mathbf{B}$ of syllable juncture and the prosodic state $\mathbf{P}$ of syllable. As shown in Fig. 1, the four prosodic constituents are delimited by seven break types denoted as $B0$, $B1$, $B2$-1, $B2$-2, $B2$-3, $B3$, and $B4$ [21]. Here, $B0$ and $B1$ are non-breaks representing the reduced and normal syllable boundaries within a PW; $B2$-1, $B2$-2 and $B2$-3 are breaks representing PW boundaries with F0 reset, short pause and pre-boundary syllable duration lengthening, respectively; $B3$ is perceived as a clear pause to represent PPh boundary; and $B4$ is defined for a breathing pause or a complete speech paragraph end. $\mathbf{P}$ is used to specify the prosodic-acoustic feature patterns of prosodic constituents. In [21], an analysis was performed to illustrate the pitch patterns of PW, PPh, and BG/PG using the affecting pattern (AP) sequences of pitch prosodic states. Affecting pattern is a scalar or vector representing the influential value or pattern from a specific affecting factor on a prosodic-acoustic feature. Three types of prosodic states, $p_n$, $q_n$ and $r_n$, are used for syllable pitch contour, duration and energy level, respectively. Thus, the complete prosodic tag sequence is $\mathbf{T} = \{\mathbf{B}, \mathbf{P}\}$, where $\mathbf{B} = \{B_n\}$ and $\mathbf{P} = \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$ are sequences of these prosodic tags defined above.

The linguistic features $\mathbf{L}$ involved in the HPM are classified into two classes. One is composed of low-level syllable-related features including lexical tone sequence $\mathbf{t} = \{t_n\}$, base-syllable sequence $\mathbf{s} = \{s_n\}$, and *final* type sequence $\mathbf{f} = \{f_n\}$. Another comprises word-level features including word length sequence $\mathbf{WL}$, part-of-speech sequence $\mathbf{POS}$, and punctuation mark sequence $\mathbf{PM}$. So, $\mathbf{L} = \{L_n\} = \{\mathbf{t}, \mathbf{s}, \mathbf{f}, \mathbf{WL}, \mathbf{POS}, \mathbf{PM}\}$.

The HPM model is formulated by

$$\begin{aligned} P(\mathbf{T}, \mathbf{A}|\mathbf{L}) &= P(\mathbf{A}|\mathbf{T}, \mathbf{L}) P(\mathbf{T}|\mathbf{L}) \\ &= P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|\mathbf{B}, \mathbf{P}, \mathbf{L}) P(\mathbf{B}, \mathbf{P}|\mathbf{L}) \\ &\approx P(\mathbf{X}|\mathbf{B}, \mathbf{P}, \mathbf{L}) P(\mathbf{Y}, \mathbf{Z}|\mathbf{B}, \mathbf{L}) P(\mathbf{P}|\mathbf{B}) P(\mathbf{B}|\mathbf{L}) \end{aligned} \tag{1}$$

where $P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L})$ is the syllable prosodic-acoustic model which describes the influences of the two types of prosodic tags and the contextual linguistic features on the variations of syllable F0 contour, duration and energy level; $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$ is the syllable-juncture prosodic-acoustic model describing the inter-syllable acoustic characteristics specified for different break type and surrounding linguistic features; $P(\mathbf{P}|\mathbf{B})$ is the prosodic state model describing the variation of prosodic state conditioned on the neighboring break type; and $P(\mathbf{B}|\mathbf{L})$ is the break-syntax model describing the dependence of break occurrence frequency on the surrounding linguistic features. In the above formulation, some assumptions are made to let the model be simple and tractable. First, the syllable-based features $\mathbf{X}$ and the two juncture-based features $\mathbf{Y}$ and $\mathbf{Z}$ are assumed to be independent so that the influences of prosodic tags and contextual features on them are separately considered. Second, the two juncture-based features $\mathbf{Y}$ and $\mathbf{Z}$ are mainly influenced by the juncture-based tags $\mathbf{B}$ and contextual feature $\mathbf{L}$. So we let them be independent of the syllable-based tags $\mathbf{P}$. Last, $\mathbf{P}$ is assumed to be independent of $\mathbf{L}$. This is a strategy used in the prosody modeling. It is motivated by the fact that the use of a simple prosodic state model $P(\mathbf{P}|\mathbf{B})$ has almost no harm to the labeling of $\mathbf{P}$ in the prosody modeling because of the availability of the prosodic-acoustic features $\mathbf{X}$. This can let the prosodic state labeling rely more on the prosodic-acoustic features. After training, we can refine the prosodic state model using the prosody-labeled training dataset when it is needed. A practice was realized in a previous study to extract the syllable pitch-level patterns of prosodic constituents of PW, PPh, and BG/PG from the $\mathbf{P}$-labeled training dataset [21]. In this study, we will create an additional model $P(\mathbf{P}|\mathbf{L})$ to describe the relation between $\mathbf{P}$ and $\mathbf{L}$ from the $\mathbf{P}$-labeled training dataset in the application to SR-controlled TTS to assist in the prediction of $\mathbf{P}$ from $\mathbf{L}$ and $\mathbf{B}$ (to be discussed in Section IV).

$P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L})$ is further divided into three sub-models for $\mathbf{sp}_n$, $sd_n$ and $se_n$:

$$\begin{aligned} P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L}) &\approx P(\mathbf{sp}|\mathbf{B},\mathbf{p},\mathbf{t})P(\mathbf{sd}|\mathbf{B},\mathbf{q},\mathbf{t},\mathbf{s}) \\ &\quad \times P(\mathbf{se}|\mathbf{B},\mathbf{r},\mathbf{t},\mathbf{f}) \\ &\approx \prod_{n=1}^{N} P\left(\mathbf{sp}_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1}\right) \\ &\quad \times P(sd_n|q_n,s_n,t_n)P(se_n|r_n,f_n,t_n) \end{aligned} \quad (2)$$

where $B_{n-1}^{n}=(B_{n-1},B_n)$; and $t_{n-1}^{n+1}=(t_{n-1},t_n,t_{n+1})$. The sub-model $P(\mathbf{sp}_n|B_{n-1}^{n},p_n,t_{n-1}^{n+1})$ is further elaborated to consider four major affecting factors and formulated as a multi-dimensional linear regression problem by

$$\mathbf{sp}_n=\mathbf{sp}_n^r+\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}_{B_{n-1},tp_{n-1}}^{f}+\boldsymbol{\beta}_{B_n,tp_n}^{b}+\boldsymbol{\mu}_{sp} \quad (3)$$

where $\mathbf{sp}_n$ is the observed log-F0 contour of syllable $n$; $\mathbf{sp}_n^r$ is the modeling residue; $\boldsymbol{\beta}_{t_n}$ and $\boldsymbol{\beta}_{p_n}$ are the affecting patterns (APs) for the affecting factors (AFs) $t_n$ and $p_n$, respectively; $tp_n$ represents the tone pair $t_n^{n+1}$; $\boldsymbol{\beta}_{B_{n-1},tp_{n-1}}^{f}$ and $\boldsymbol{\beta}_{B_n,tp_n}^{b}$ are the forward and backward coarticulation APs contributed from syllable $n-1$ and syllable $n+1$, respectively; and $\boldsymbol{\mu}_{sp}$ is the global mean of pitch vector. By assuming that $\mathbf{sp}_n^r$ is zero-mean and normally distributed, i.e., $N(\mathbf{sp}_n^r;0,R_{sp})$, we have

$$\begin{aligned} &P\left(\mathbf{sp}_n|p_n,B_{n-1}^{n},t_{n-1}^{n+1}\right) \\ &=N(\mathbf{sp}_n;\boldsymbol{\beta}_{t_n}+\boldsymbol{\beta}_{p_n}+\boldsymbol{\beta}_{B_{n-1},tp_{n-1}}^{f}+\boldsymbol{\beta}_{B_n,tp_n}^{b}+\boldsymbol{\mu}_{sp},R_{sp}) \end{aligned} \quad (4)$$

Similarly, the other two sub-models are formulated by

$$P(sd_n|q_n,s_n,t_n)=N(sd_n;\gamma_{t_n}+\gamma_{s_n}+\gamma_{q_n}+\mu_{sd},R_{sd}) \quad (5)$$

$$P(se_n|r_n,f_n,t_n)=N(se_n;\omega_{t_n}+\omega_{f_n}+\omega_{r_n}+\mu_{se},R_{se}) \quad (6)$$

where $\gamma'$s and $\omega'$s represent APs of syllable duration and syllable energy level, respectively; $\mu_{sd}$ and $\mu_{se}$ are their global means; and $R_{sd}$ and $R_{se}$ are variances of modeling residues.

$P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L})$, is further divided into five sub-models by

$$\begin{aligned} P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L}) &\approx P(\mathbf{pd},\mathbf{ed},\mathbf{pj},\mathbf{dl},\mathbf{df}|\mathbf{B},\mathbf{L}) \\ &\approx \prod_{n=1}^{N-1} P(pd_n,ed_n,pj_n,dl_n,df_n|B_n,L_n) \\ &\approx \prod_{n=1}^{N-1} \{G(pd_n;\alpha_{B_n,L_n},\beta_{B_n,L_n}) \\ &\quad \times N(ed_n;\mu_{ed,B_n,L_n},\sigma_{ed,B_n,L_n}^2) \\ &\quad \times N(pj_n;\mu_{pj,B_n,L_n},\sigma_{pj,B_n,L_n}^2) \\ &\quad \times N(dl_n;\mu_{dl,B_n,L_n},\sigma_{dl,B_n,L_n}^2) \\ &\quad \times N(df_n;\mu_{df,B_n,L_n},\sigma_{df,B_n,L_n}^2)\} \end{aligned} \quad (7)$$

where $G(pd_n;\alpha_{B_n,L_n},\beta_{B_n,L_n})$ is a Gamma distribution for $pd_n$; and the other four features are all modeled as normal distributions. Since the space of $L_n$ is large, the CART algorithm [24] with the node splitting criterion of maximum likelihood (ML) gain with a minimum sample size constraint is adopted to concurrently classify the five features for each break type according to a question set.

$P(\mathbf{P}|\mathbf{B})$ is further divided into three sub-models by

$$\begin{aligned} P(\mathbf{P}|\mathbf{B}) &\approx P(\mathbf{p}|\mathbf{B})P(\mathbf{q}|\mathbf{B})P(\mathbf{r}|\mathbf{B}) \\ &\approx P(p_1)P(q_1)P(r_1) \\ &\quad \cdot \prod_{n=2}^{N} [P(p_n|p_{n-1},B_{n-1})P(q_n|q_{n-1},B_{n-1}) \\ &\quad \times P(r_n|r_{n-1},B_{n-1})] \end{aligned} \quad (8)$$

Lastly, the break-syntax model $P(\mathbf{B}|\mathbf{L})$ is approximated by

$$P(\mathbf{B}|\mathbf{L}) \approx \prod_{n=1}^{N-1} P(B_n|L_n) \quad (9)$$

where $P(B_n|L_n)$ is the break type model for juncture $n$. We also realize $P(B_n|L_n)$ by the CART algorithm.

The HPM is trained automatically from a prosody-unlabeled speech corpus by the PLM algorithm [21] which is a sequential optimization procedure based on the ML criterion to jointly label the prosodic tags for all utterances in the training corpus and estimate the parameters of all 12 prosodic sub-models. It

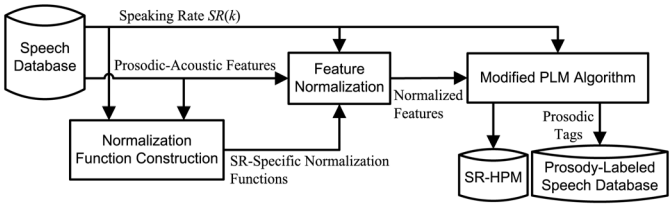



Fig. 2. A schematic diagram of the proposed speaking rate modeling approach.

first defines an objective likelihood function $Q$ formed by these 12 sub-models for all utterances, and then performs a multi-step iterative procedure to re-label the prosodic tags of each utterance with the goal of maximizing $Q$ and to update the parameters of all prosodic models sequentially and iteratively.

## III. The Speaking Rate Modeling

Fig. 2 shows a schematic diagram of the proposed speaking rate modeling method. For each utterance $k$, the average number of syllables per second calculated with all pauses being excluded is taken as a measure of speaking rate and denoted as $SR(k)$. Note that $SR(k)$ is also known as articulation rate. The prosodic-acoustic features of the utterance are then normalized by SR-specific normalization functions to compensate the influences of speaking rate on them. Those SR-specific normalization functions are constructed using the statistics of the prosodic-acoustic features of all utterances in the whole database. Lastly, a modified version of the PLM algorithm [21] is employed to construct the SR-HPM and label prosodic tags of all utterances, simultaneously. The modification of the PLM algorithm lies in letting some model parameters be dependent variables of $SR(k)$. In the following subsections, we describe the method in detail.

### A. The Speech Database

A speech database containing four parallel speech corpora of a female professional announcer with fast, normal, medium and slow speaking rates is used in the speaking rate modeling. The associated texts of each dataset contain 380 short paragraphs selected from the Sinica Treebank Version 3.0 [25]. Each paragraph is composed of several sentences. Originally, the database contains, in total, 1,520 utterances with 208,768 syllables. After excluding utterances of bad recording quality, 1,478 utterances with 203,746 syllables are used in the study. All utterances are segmented into syllables, and then syllable pitch contours are found. Fig. 3 shows the histogram (utterance count) of speaking rate of the database. As shown in the figure, the *SR*s of utterances in these four speech corpora distribute widely in the range of 3.4-6.8 syl/sec and overlapped seriously. The database is divided into a training set with 183,795 syllables and a test set with 19,951 syllables. The training set is used to construct the SR-specific feature normalization functions and the SR-HPM model, while the test set is used for outside test.

### B. Prosodic-Acoustic Feature Normalization

The prosodic-acoustic feature normalization is performed in the preprocessing stage of the proposed speaking rate modeling to equalize the influences of different speaking rates on prosodic-acoustic features so that we can, in the following

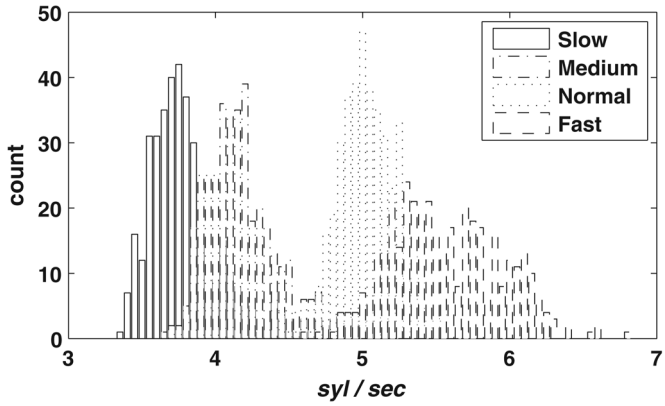



Fig. 3. Histogram of utterance's speaking rate of four databases used in the study.

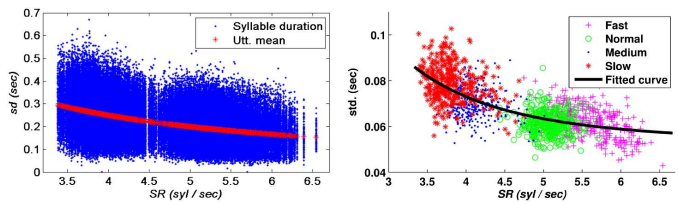



Fig. 4. The scatter plots of $sd_{n,k}$ and utterance-wise mean vs. $SR(k)$ (left) and utterance-wise standard deviation vs. $SR(k)$ (right).

stages, first build an HPM model with parameters in common for all speaking rates, and then make some parameters of the HPM be SR-dependent. A popular feature normalization method is the z-score normalization using the utterance-based mean and standard deviation of the processing feature. Although the method is simple and effective, it has a problem in the speaking rate-controlled TTS application to select a proper denormalization function solely from *SR*. This can be justified based on the fact that many utterances in our database have similar *SR*s but with quite different means and standard deviations. As the conventional z-score normalization is applied, it will be a problem to choose proper mean and standard deviation (which are unknown) from *SR* for denormalization. A possible way to solve the problem is using the local averages of mean and standard deviation. But this will result in an inconsistency between normalization and denormalization. To avoid the drawback, we take care of each prosodic-acoustic feature separately by designing smooth normalization functions from the statistics of the feature of all utterances in the database. In the following subsections, we discuss the normalizations of syllable duration, syllable-juncture pause duration, syllable pitch contour, and syllable energy level in detail.

*1) Syllable Duration Normalization:* Fig. 4 displays the scatter plots of syllable duration $sd_{n,k}$ and utterance-wise mean vs. $SR(k)$ (left) and utterance-wise standard deviation vs. $SR(k)$ (right). Here, $n$ and $k$ denote indices for syllable and utterance. From the left panel, it is observed that $sd_{n,k}$ scatters over a larger range for smaller $SR(k)$ and the utterance-wise syllable duration mean is inversely proportional to $SR(k)$. From the right panel, we find that the scattering of utterance-wise standard deviation also depends on $SR(k)$ with a trend of decreasing as $SR(k)$. We also find that utterance-wise standard deviation can be quite different even when their $SR(k)$ are very close. As

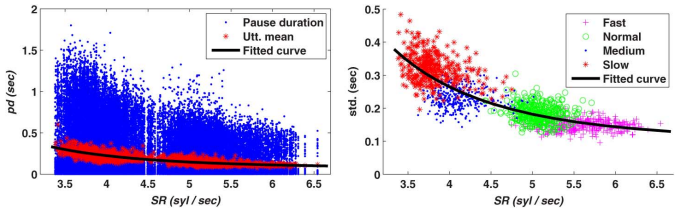



Fig. 5. The scatter plots of $pd_{n,k}$ and utterance-wise mean vs. $SR(k)$ (left) and utterance-wise standard deviation vs. $SR(k)$ (right).

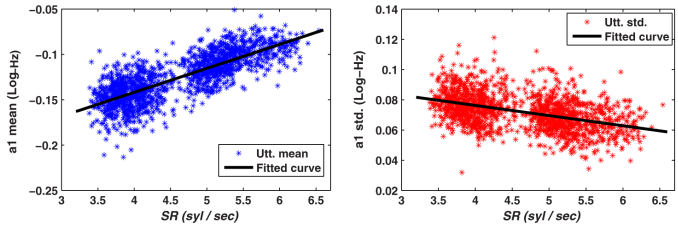



Fig. 6. The scatter plots and fitting lines for the utterance-wise mean and standard deviation of $\mathbf{sp}_{n,k}(2)(=a^1_{n,k})$ with tone 4.

discussed before, directly using the utterance-wise mean and standard deviation for syllable duration normalization is hence improper. In this study, a smooth function of *SR*, constructed via fitting the scatter plot of utterance-wise standard deviation with a second-order polynomial, is employed to generate a smoothed standard deviation for each utterance to replace the original one for syllable duration normalization. Since utterance-wise standard deviation is in a unit of time, it is more suitable to take the utterance-wise syllable duration mean instead of $SR(k)$ as the independent variable in the syllable duration normalization function. For modeling convenience, we define a new variable $x(k)$ as $x(k) = 1/SR(k)$. The normalization for $sd_{n,k}$ is then formulated by

$$sd'_{n,k} = (sd_{n,k} - x(k))/\tilde{\sigma}^{sd}(x(k)) \cdot \sigma_g^{sd} + \mu_g^{sd} \tag{10}$$

where

$$\tilde{\sigma}^{sd}(x(k)) = a_1(x(k))^2 + b_1 \cdot x(k) + c_1 \tag{11}$$

is the least-square fitted smooth curve for standard deviation; $sd'_{n,k}$ is the SR-normalized version of syllable duration; and $\mu_g^{sd}$ and $\sigma_g^{sd}$ are the global mean and standard deviation of syllable duration of the whole database.

*2) Pause Duration Normalization:* Syllable-juncture pause duration $pd_{n,k}$ can be deviated wildly from very small values for non-break junctures, to medium values for minor-break junctures, and to very large values for major-break junctures. Since the influences of speaking rate on pause duration are more serious for both minor and major breaks, statistics of medium and long pause durations are more emphasized in constructing the pause duration normalization function. This is realized via calculating the statistics of pause duration using only data with values larger than 5 ms. The dominating effect of very short pause durations due to their large total amount labeled for most intra-word syllable junctures and some inter-word junctures can therefore be avoided. Note that the normalization operations are performed on all pause durations. This is to keep the normalized values of small pause durations in the same order as their original counterparts. Fig. 5 displays the scatter plots of $pd_{n,k}$ and utterance-wise pause-duration mean vs. $SR(k)$ (left) and utterance-wise standard deviation vs. $SR(k)$ (right). It can be found from the figure that both utterance-wise pause-duration mean and standard deviation scatter to larger-valued ranges as $SR(k)$ decreases. Based on the same idea of treating $\sigma^{sd}(x(k))$, two second-order polynomials are employed to fit the trends of scatterings of utterance-wise pause-duration mean and standard deviation by

$$\tilde{\mu}^{pd}(x(k)) = a_2(x(k))^2 + b_2 \cdot x(k) + c_2 \tag{12}$$

$$\tilde{\sigma}^{pd}(x(k)) = a_3(x(k))^2 + b_3 \cdot x(k) + c_3 \tag{13}$$

The results of curve-fitting are displayed in Fig. 5.

The implementation of pause duration normalization is discussed as follows. In this study, $pd_{n,k}$ is modeled as a Gamma distribution $G(pd_{n,k}; \alpha^{pd}(x(k)), \beta^{pd}(x(k)))$ with two parameters $\alpha^{pd}(x(k))$ and $\beta^{pd}(x(k))$. A Gamma distribution-normalization scheme is adopted to normalize cumulative distribution function (*cdf*) of pause duration (of a specified *SR*) to a reference one. This is realized by firstly calculating the smoothed $\tilde{\alpha}^{pd}(x(k))$ and $\tilde{\beta}^{pd}(x(k))$ from $\tilde{\mu}^{pd}(x(k))$ and $\tilde{\sigma}^{pd}(x(k))$ by

$$\tilde{\alpha}^{pd}(x(k)) = (\tilde{\mu}^{pd}(x(k)))^2/(\tilde{\sigma}^{pd}(x(k)))^2 \tag{14}$$

$$\tilde{\beta}^{pd}(x(k)) = (\tilde{\sigma}^{pd}(x(k)))^2/\tilde{\mu}^{pd}(x(k)) \tag{15}$$

Then, an SR-specific normalization function is derived and used to normalize $pd_{n,k}$ by

$$pd'_{n,k} = G^{-1}(G(pd_{n,k}; \tilde{\alpha}^{pd}(x(k)), \tilde{\beta}^{pd}(x(k)));\ \alpha_g^{pd},\ \beta_g^{pd}) \tag{16}$$

where $G^{-1}$ is the inverse function of $G$; $\alpha_g^{pd}$ and $\beta_g^{pd}$ are parameters of the reference Gamma distribution calculated from global mean and standard deviation.

*3) Syllable log-F0 Contour Normalization:* The *log*-F0 contour of syllable $n$ in utterance $k$ is represented by $\mathbf{sp}_{n,k} = [\,a^0_{n,k} \quad a^1_{n,k} \quad a^2_{n,k} \quad a^3_{n,k}\,]^T$ with components representing respectively the mean, slope, acceleration and curvature of the contour [22]. By observing the scatter plots of the utterance-wise mean and standard deviation of each coefficient vs. $SR(k)$, we find that it is necessary to compensate the effects of speaking rate on syllable *log*-F0 contour for each component of each lexical tone, i.e.,

$$\mathbf{sp}'_{n,k}(i) = \frac{\mathbf{sp}_{n,k}(i) - \tilde{\mu}^{sp}(x(k), t_n, i)}{\tilde{\sigma}^{sp}(x(k), t_n, i)}\sigma_g^{sp}(t_n, i) + \mu_g^{sp}(t_n, i) \quad \text{for } t_n = 1 \sim 5 \text{ and } i = 1 \sim 4 \tag{17}$$

is the SR-specific normalization functions for the $i$-th component of $\mathbf{sp}_{n,k}$ with tone $t_n$, where $\tilde{\mu}^{sp}(x(k), t_n, i)$ and $\tilde{\sigma}^{sp}(x(k), t_n, i)$ represent respectively the smoothed mean and standard deviation; and $\mu_g^{sp}(t_n, i)$ and $\sigma_g^{sp}(t_n, i)$ are the $i$-th components of the global mean and standard deviation vectors for tone $t_n$. It is noted that the five tones in Mandarin are high-level, middle-rising, low-dipping, high-falling and unstressed low-level tones, and are commonly referred to as tone 1 - tone 5. Two first-order polynomials of $x(k)$ are constructed for $\tilde{\mu}^{sp}(x(k), t_n, i)$ and $\tilde{\sigma}^{sp}(x(k), t_n, i)$, respectively. Fig. 6 shows an example of the scatter plots and fitted curves for the mean and standard deviation of $\mathbf{sp}_n(2)$ (i.e., slope $a^1_n$) with tone 4 (high-falling tone). It can be seen from these two figures

that both mean and standard deviation are correlated with $SR(k)$. The slope of tone 4 becomes more negative as $SR(k)$ becomes smaller (slower speech). This mainly results from the linear normalization of the orthogonal expansion operation in feature extraction [22], and may also result in part from the more complete pronunciation of tone 4 for slow speech.

*4) Syllable Energy Level Normalization:* Generally, syllable energy level is highly correlated with the recording conditions, e.g. distance between microphone and speaker, the gain of the microphone set on the recording equipment, etc. By observing the scatter plots of syllable energy level, we find that they truly depend on the recording condition of each utterance but not speaking rate. Therefore, we simply let them be z-score normalized to the global mean and standard deviation on an utterance-by-utterance basis.

### C. The Modified PLM Method

After feature normalization, a modified version of the PLM algorithm proposed previously [21] is employed to automatically train an SR-HPM model and label all utterances with the two types of prosodic tags: prosodic state and break type. The modification of the PLM algorithm mainly lies in letting some parameters of the HPM be SR-dependent to account for the influences of speaking rate. Since the occurrence frequency of break is known to highly depend on the speaking rate [10], we consider it by letting the break-syntax model, which is a decision tree describing the relation between the occurrence frequencies of 7 break types and various contextual linguistic features, be SR-dependent. Besides, we also let the parameters of two prosodic state sub-models for syllable pitch and duration be SR-dependent.

The modified PLM algorithm is formulated in the same way like the original one [21] as a sequential optimization problem except that the break-syntax model and the two prosodic state sub-models in the $Q$ function becomes SR-dependent. In this study, the SR-dependent break-syntax model $P(B_{n,k}|L_{n,k},x(k))$ is constructed by two steps. Note that the subscript $k$ additionally added is to specify utterance. In Step 1, the marginal probability $P(B_{n,k}|L_{n,k})$ is firstly estimated from the labeled $B_{n,k}$ by the CART as in the original PLM algorithm [21]. In Step 2, the scatter plot of the occurrence frequency vs. $x(k)$ for each break type in each leaf node of the decision tree constructed in Step 1 is formed, and then linearly fitted and normalized to obtain $P(B_{n,k}|L_{n,k},x(k))$, i.e.,

$$\begin{aligned} P(B_{n,k}=m|L_{n,k}=j,x(k)) &= \frac{P'(B_{n,k}=m|L_{n,k}=j,x(k))}{\sum_{m'\in \text{ all break types}} P'(B_{n,k}=m'|L_{n,k}=j,x(k))} \\ &\approx \frac{c_{m,j}x(k)+d_{m,j}}{\sum_{m'\in\ all\ break\ types} c_{m',j}x(k)+d_{m',j}} \end{aligned} \tag{18}$$

where $j$ denotes the index of the leaf node associated with the linguistic features $L_{n,k}$; and $c_{m,j}$ and $d_{m,j}$ are linear regression coefficients estimated from the histogram of break type $m$ in leaf node $j$. The two SR-dependent prosodic state sub-models for syllable pitch contour and duration are constructed by a bin-based normalization scheme, i.e., (see equation (19) shown at the bottom of the page), where $ps\in\{p,q\}$ is a type of prosodic state; and $bin(x(k))$ is a bin of the histogram of $x(k)$ for the triplet $(i,j,l)$.

In summary, the following objective likelihood function is used in the modified PLM (see equation (20) shown at the bottom of the page), where $K$ is the total number of utterances in the training set.

### D. Experimental Results of Speaking Rate Modeling

The performance of the proposed speaking rate modeling scheme was examined via investigating the modeling er-

$$\begin{aligned} P(ps_{n,k}=i|ps_{n-1,k}=j,B_{n-1,k}=l,x(k)) &= \frac{P'(ps_{n,k}=i|ps_{n-1,k}=j,B_{n-1,k}=l,bin(x(k)))}{\sum_{i'\in \text{ all states}} P'(ps_{n,k}=i'|ps_{n-1,k}=j,B_{n-1,k}=l,bin(x(k)))} \end{aligned} \tag{19}$$

$$\begin{aligned} Q=\prod_{k=1}^{K}\Bigg\{ &\left(\prod_{n=1}^{N-1} P(B_{n,k}|L_{n,k},x(k))\right) \\ &\left(\begin{array}{l} P(p_{1,k}|x(k))P(q_{1,k}|x(k))P(r_{1,k}) \\ \prod_{n=2}^{N}\left[\begin{array}{l} P(p_{n,k}|p_{n-1,k},B_{n-1,k},x(k))P(q_{n,k}|q_{n-1,k},B_{n-1,k},x(k)) \\ P(r_{n,k}|r_{n-1,k},B_{n-1,k}) \end{array}\right] \end{array}\right) \\ &\left(\prod_{n=1}^{N}\left[\begin{array}{l} P(\mathbf{sp}'_{n,k}|p_{n,k},B_{n-1,k}^{n,k},t_{n-1,k}^{n+1,k})P(sd'_{n,k}|q_{n,k},s_{n,k},t_{n,k}) \\ P(se'_{n,k}|r_{n,k},f_{n,k},t_{n,k}) \end{array}\right]\right) \\ &\left(\begin{array}{l} \prod_{n=1}^{N-1}\Big[G(pd'_{n,k};\alpha_{B_{n,k},L_{n,k}},\beta_{B_{n,k},L_{n,k}})N(ed_{n,k};\mu_{ed,B_{n,k},L_{n,k}},\sigma^2_{ed,B_{n,k},L_{n,k}}) \\ \cdot N(pj_{n,k};\mu_{pj,B_{n,k},L_{n,k}},\sigma^2_{pj,B_{n,k},L_{n,k}})N(dl_{n,k};\mu_{dl,B_{n,k},L_{n,k}},\sigma^2_{dl,B_{n,k},L_{n,k}}) \\ \cdot N(df_{n,k};\mu_{df,B_{n,k},L_{n,k}},\sigma^2_{df,B_{n,k},L_{n,k}})\Big] \end{array}\right)\Bigg\} \end{aligned} \tag{20}$$

TABLE I
MODELING ERROR BY THE PROPOSED, HPM AND Z-SCORE NORMALIZATION METHODS: (A) VARIANCES FOR SYLLABLE PITCH CONTOUR, DURATION AND ENERGY LEVEL; (B) RMSE (*ms*) OF RECONSTRUCTED PAUSE DURATION. I:INSIDE TEST USING TRAINING SET; O: OUTSIDE TEST USING TEST SET.

| | | **sp** (x$10^{-4}$ (log-*Hz*)$^2$) | *sd* (*ms*$^2$) | *se* (*dB*$^2$) |
|---|---|---|---|---|
| Original | I | [488, 81, 16, 4] | 6000 | 38.85 |
| | O | [528, 136, 19, 5] | 5700 | 47.46 |
| Proposed | I | [4.2, 34, 11.3, 3.7] | 67 | 0.53 |
| | O | [4.3, 35, 11.9, 3.8] | 73 | 0.57 |
| HPM | I | [4.6, 35, 11.4, 3.7] | 120 | 1.26 |
| | O | [4.7, 38, 12.3, 3.8] | 143 | 1.64 |
| z-score | I | [18, 35, 11.2, 3.7] | 152 | 0.68 |
| | O | [16, 37, 12.4, 3.8] | 129 | 0.79 |

(a)

UNIT: *ms*

| Break type | | *B*0 | *B*1 | *B*2-1 | *B*2-2 | *B*2-3 | *B*3 | *B*4 | total |
|---|---|---|---|---|---|---|---|---|---|
| Proposed | I | 2.4 | 18.5 | 24.9 | 86.3 | 30.8 | 101 | 149 | 51.4 |
| | O | 5.8 | 18.6 | 25.0 | 85.1 | 28.4 | 93.8 | 140 | 53.9 |
| HPM | I | 1.3 | 18.0 | 26.1 | 109 | 31.9 | 143 | 219 | 67.1 |
| | O | 4.5 | 17.6 | 22.2 | 99.6 | 27.8 | 122 | 203 | 72.8 |
| z-score | I | 14.6 | 22.1 | 17.3 | 45.2 | 25.0 | 128 | 159 | 60.4 |
| | O | 9.7 | 17.4 | 13.2 | 65.1 | 18.0 | 123 | 172 | 67.0 |

(b)

rors, the parameters of the trained SR-HPM model and the prosody labeling results on the speech database discussed in Subsection III-A. The modified PLM algorithm took 94 iterations to reach a convergence.

Table I lists the modeling errors of syllable pitch contour, syllable duration, syllable energy level, and syllable-juncture pause duration for the inside test (I) using training set and the outside test (O) using test set by the proposed method and the two comparing methods: the original HPM without SR normalization (HPM) and the SR-HPM using utterance-based z-score normalization (z-score). Here local averages of mean and standard deviation are used for z-score denormalization. As shown in Table I(A), the variances of modeling residuals of the three syllable-based prosodic-acoustic features became very small for both inside and outside tests by the proposed method as compared with the variances of the original unnormalized features (Original). This showed the effectiveness of the modeling scheme by the proposed SR-HPM method. We also find from the table that the proposed method outperformed both the HPM and z-score methods. Without considering the effect of speaking rate, the modeling efficiencies of the HPM method were degraded on pitch mean and syllable energy level, and largely on syllable duration. It is a surprise that the modeling error of syllable energy level was smaller for SR-HPM which didn't consider the effect of SR on its feature normalization. This might result from the better modeling of break in the SR-HPM. For the z-score method, the modeling efficiencies degraded seriously on pitch mean and syllable duration. This mainly resulted from the inconsistencies on the feature normalizations and denormalizations. Table I(B) displays the RMSEs of the reconstructed pause duration for different break types. It is noted that the data counts of these seven break types are different for all cases. It can be found from Table I(B) that RMSEs were large only for $B2$-2, $B3$ and $B4$ for the proposed SR-HPM method. Since the pause durations of these three break types were inherently longer with much larger dynamic ranges, these results were reasonable. We also find from the last column of Table I(B) that the overall performance of the proposed method was better than these of the HPM and z-score methods. The performance of HPM degraded seriously for $B2$-2, $B3$ and $B4$ because of the wide spreading of the original pause durations without performing SR-normalization. Actually, many fast-speech utterances in the normal and fast corpora had no syllable junctures being labelled as $B4$ by the HPM method. The z-score method had large RMSEs for $B3$ and $B4$ resulting from the inconsistencies of the pause duration normalizations and denormalizations. Based on above discussions, we concluded that the proposed SR-HPM method outperformed both the conventional HPM method and the z-score normalization method. In the following, we analyzed the parameters of the SR-HPM in more detail.

TABLE II
TREs OF SYLLABLE PITCH CONTOUR, DURATION, AND ENERGY LEVEL MODELING FOR DIFFERENT AGGREGATION OF APs

| **sp** | | *sd* | | *se* | |
|---|---|---|---|---|---|
| APs | TRE | Aps | TRE | APs | TRE |
| tone *t* | 67.3% | tone *t* | 70.6% | tone *t* | 61.4% |
| + coarticu. | 63.2% | + syllable *s* | 50.1% | + *final f* | 48.0% |
| + prosodic state *p* | 0.8% | + prosodic state *q* | 1.4% | + prosodic state *r* | 1.9% |

Table II displays the total residual errors (TREs) of the SR-HPM modeling for the three reconstructed syllable-based prosodic-acoustic features using different combinations of affecting patterns (APs) with denormalization. Here, TRE is defined as the ratio of the variance of the modeling residual with respect to that of the original unnormalized feature [21], [23]. Values in the table show the effects of removing the influences of APs of some affecting factors considered on $\mathbf{sp}$, $sd$, and $se$. These results generally agreed with those achieved in our previous study of building individual HPMs for the four parallel speech corpora [10].

Fig. 7 displays the denormalized values of means of pause duration at the root nodes of 7 break-type decision trees vs. $SR(k)$. They were averaged values to show the overall trend with respect to speaking rate. Those values could be further refined by tracing down their corresponding decision trees as the contextual linguistic feature $L_n$ was considered. It is found from the figure that $B2$-2, $B3$ and $B4$ had significantly large pause durations which increased nonlinearly as $SR(k)$ decreased. These values matched well with the results of 4 individually-trained HPMs (shown as $*$, o and $\Delta$) [10] and agreed well with our prior knowledge about break's pause duration [1], [26].

Fig. 8 displays the basic tone patterns (after denormalization) of the SR-HPM for two typical speaking rates of fast ($SR = 5.6$ syl/sec) and slow ($SR = 3.7$ syl/sec) speeches. This demonstrated how the SR-HPM shrank and expanded the five tone patterns for fast and slow speeches. As shown in the

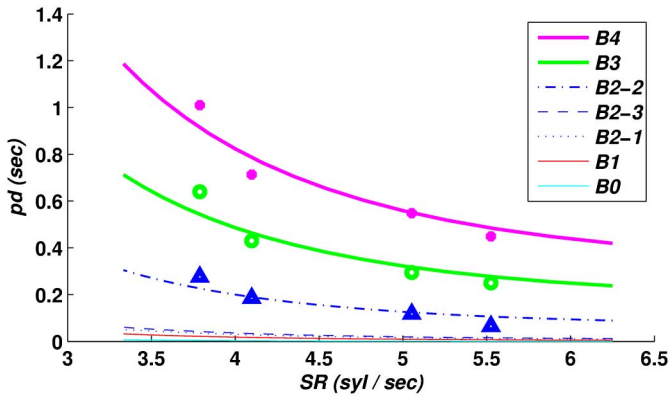



Fig. 7. Average pause durations of 7 break types of SR-HPM vs. $x(k)$. Here, $*$, o and $\triangle$ denotes the values in the HPMs individually trained from 4 databases.

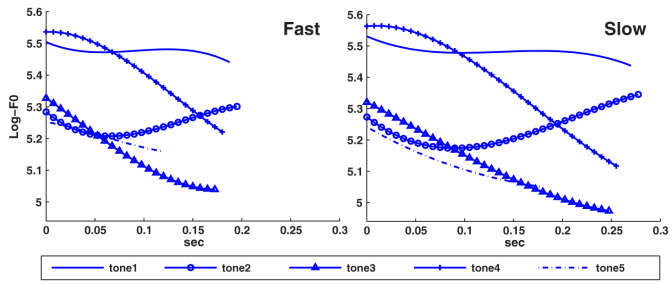



Fig. 8. Basic tone patterns (after denormalization) of SR-HPM for two speaking rates of fast ($SR = 5.6$ syl/sec) and slow ($SR = 3.7$ syl/sec) speech.

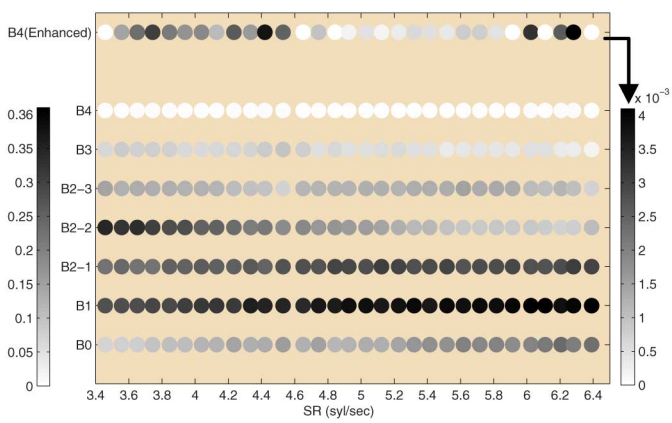



Fig. 9. Distributions of 7 break types labeled for all non-PM inter-word syllable junctures. Darker nodes represent higher probabilities.

figure, all five pitch contours of the slow speech were longer and spanned to larger ranges as compared with their counterparts of the fast speech. If we normalized the two pitch contours of the two speaking rates with the same tone (e.g. tone 4) to the same length, the one of the fast speech becomes more flat. This showed that pitch contours of faster speech were pronounced shorter with similar slope or more flat with similar length. We note that these patterns would be further refined by aggregating other APs as more affecting factors were considered.

Fig. 9 shows the distributions of the break types labeled for all non-PM inter-word syllable junctures. It is found from the figure that they were largely labeled as non-pause breaks (i.e., $B0$, $B1$, $B2$-1) for fast speech; while more short- to long-pause

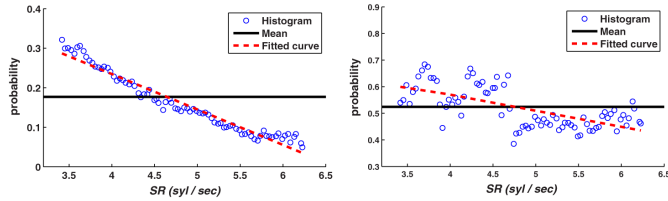



Fig. 10. Two examples of occurrence probability of break type: (left) $B2$-2 in a non-PM inter-word node and (right) $B4$ in a PM node.

| |
|---|
| 依據(According to) 行政院(Executive Yuan) 主計處(Directorate-General of Budget, Accounting and Statistics) 的(*an possessive marker*, 's) 統計(statistics)，十月份(October) 一(first) 到(to) 二十日(twentieth) ， 我國(our country) 出口(export) 及(and) 進口(import) 金額(amount of money) 比起(compare with) 去年(last year) 同期(the same time) 均有(all) 增加(increase)， |
| 依據 行政院 主計處 的 統計 @，十月份 * 一 到 二十日 / ，我國 出口 及進口 金額 / 比起 去年 同期 * 均有 增加 @， |
| 依據 行政院 主計處 的 統計 @，十月份 * 一 到 二十日 / ，我國 出口 * 及 進口 金額 / 比起 去年 同期 * 均有 增加 @， |
| 依據 * 行政院 主計處 的 統計 @，十月份 / 一 到 * 二十日 /，我國 出口 * 及 進口 金額 / 比起 去年 同期 * 均有 增加 @， |
| 依據 / 行政院 * 主計處 的 統計 @，十月份 / 一 * 到 * 二十日 @，我國 出口 * 及 進口 金額 / 比起 去年 同期 * 均有 增加 @， |

Fig. 11. An excerpted break labeling results of text-parallel utterances in different speaking rate (from second raw to bottom, fast to slow). Note that only pause-related break types, i.e. $B4$ (\@), $B3$(/) and $B2$-2($*$) are displayed.

breaks (i.e., $B2$-2, $B3$, $B4$) occurred for slow speech. This generally agreed with the prior knowledge that speakers tend to insert more breaks within a sentence as they speak slower [1], [26]. It is worthy to note that the proper break-type labeling was ascribed in part to the pause-duration normalization. Without its contribution, break-type tags will be largely down-graded to the break types of shorter pause for fast speech and up-graded to the break types of longer pause for slow speech, respectively.

Fig. 10 displays two examples of the SR dependency of the parameters of the SR-dependent break-syntax model: one for short-pause minor break $B2$-2 in a node associated with non-PM inter-word and another for long-pause major break $B4$ in a PM node. The left panel shows that the occurrence probability of $B2$-2 at the non-PM inter-word juncture increased linearly as *SR* increased. This matched well with the results of $B2$-2 labeling for non-PM inter-word syllable junctures shown in Fig. 9. The right panel shows that the occurrence probability of $B4$ at the PM juncture had high average value of 0.52 and was weakly correlated with *SR*. This was resulted from the fact that almost all major PMs (i.e., period, semicolon, colon, exclamation mark, question mark) and some commas were labeled as $B4$ regardless of the speaking rate. These findings agreed well with the prior knowledge about Mandarin speech prosody [1], [26].

Fig. 11 illustrates an example of break labeling for four parallel utterances. It shows that all PMs were labeled as $B4$ or $B3$, while more inter-word junctures were labeled as $B2$-2 or $B3$ as $SR(k)$ decreased. This agreed with the trend shown in Fig. 10.

Since the z-score normalization method is not better in SR-HPM modeling and has the problem of selecting a proper denormalization function from a given *SR* for prosody generation, we therefore do not consider using it in the following study on SR-controlled Mandarin TTS.

## IV. An Application to SR-Controlled Mandarin TTS

In this section, we use the SR-HPM to design and implement a speaking rate-controlled Mandarin TTS system. The study focuses on the prosody generation to synthesize natural speech for any input Chinese text with the speaking rate given in the range of 3.4-6.8 syl/sec.

### A. Review of Prosody Generation methods

Many existing methods of prosody generation for TTS were proposed in the past [27]–[37]. They can be roughly grouped into two categories: direct modeling methods and multi-component representation methods. A direct modeling method adopts a data-driven approach to construct mapping functions from input linguistic features to output prosodic-acoustic features [27]–[34]. Nowadays, the popular direct modeling approach is to simultaneously construct spectrum, F0 and duration model by using the HMM-based speech synthesis system (HTS) [31]–[34]. The HTS adopts the decision tree method for state duration modeling and HMM with multi-space distribution for F0 modeling. It is worthy to note that HTS can generate synthesized speech with various speaking rate via setting the $\rho$ factor in the HTS engine API [34]. A multi-component representation method superimposes several prototypical contours of multi-level prosodic constituents or syntactic units for each prosodic-acoustic feature in a hierarchical way [30], [35]–[37].

From above literature review, we claim that the proposed method is different from those previous studies on two aspects. First, none of those studies considered the method of speaking rate control other than the methods of proportional duration adjustment and changing the $\rho$ factor in HTS. Even though some studies [10], [15], [19] adopted the model interpolation method for speaking rate control, they did not sophisticatedly model the effects of speaking rate as a continuous input variable in prosody generation. Second, the construction of prosodic models in the literatures generally relied on the availability of prosody-labeled speech corpora which were manually prepared in advance. On the contrary, the proposed method uses the SR-HPM model trained automatically from a large unlabeled speech corpus by the modified PLM algorithm.

### B. The Proposed TTS System

Fig. 12 displays a block diagram of the proposed TTS system. The system consists of three parts: text analysis, prosody generation, and speech synthesis. The text analyzer used is a conditional random field-based linguistic processor specially designed to generate the linguistic features of $\mathbf{L} = \{L_n\} = \{\mathbf{t}, \mathbf{s}, \mathbf{f}, \mathbf{WL}, \mathbf{POS}, \mathbf{PM}\}$ from the given raw input Chinese text. The prosody generation is powered by the following three blocks: break type prediction, prosodic state prediction and SR-dependent prosodic feature generation. The prosodic features generated by the proposed method are the predicted syllable pitch contour $\widehat{\mathbf{sp}}_n$, syllable duration $\widehat{sd}_n$, syllable energy-level $\widehat{se}_n$ and syllable-juncture pause duration $\widehat{pd}_n$. These three blocks are driven by the sub-models of the SR-HPM trained in Section III, a refined SR-dependent break-syntax model, and a newly-added prosodic state-syntax model. The need of the latter two models is explained as follows. In the prosody modeling discussed in Section III, we train the SR-HPM and label prosodic tags simultaneously from a prosody unlabeled database with the prosodic-acoustic features available for use. A strategy of using simpler break-syntax model and prosodic state model is adopted to let the prosody labeling rely more on prosodic-acoustic features and less on these two models. In TTS application, the task is to predict prosodic tags purely from linguistic features. These two trained simpler models are not good enough for the task without the help of prosodic-acoustic features. An additional training phase is hence engaged to refine the simpler break-syntax model and train the new prosodic state-syntax model for assisting in predicting prosodic tags in the TTS application by using the prosody labeling results of the training set obtained in Subsection III-D. The speech synthesis is implemented by an HMM-based speech synthesizer. The synthesized speech is generated according to the predicted prosodic-acoustic features by the proposed prosody generation method. The details of the system are described in the following subsections.

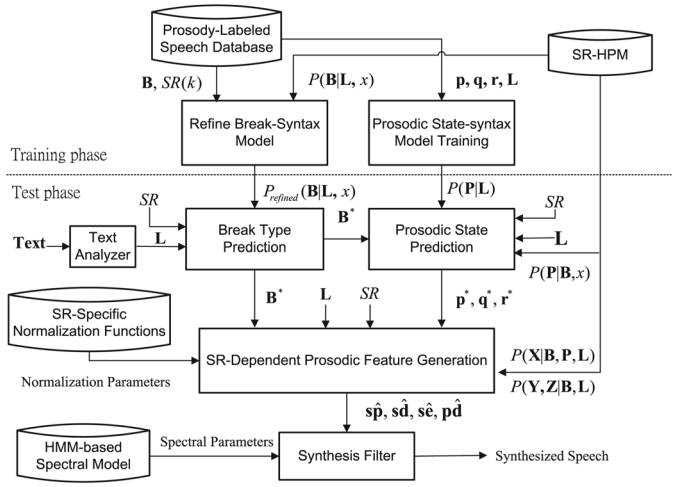



Fig. 12. A block diagram of the proposed speaking rate-controlled Mandarin TTS system.

*1) Prosody Generation:* In most prosody generation methods, break prediction is used to generate prosodic structure from given linguistic features. Many break prediction methods have been proposed in the past, including hierarchical stochastic model [38], N-gram model [39], classification and regression tree (CART) [40], [41], Markov model [42], etc. In this study, we simply apply a refined version of the SR-dependent break-syntax model to the break prediction. Recall that the SR-dependent break syntax model was trained by a specially designed training procedure based on the CART algorithm as illustrated in Subsection III-C. The proposed approach is advantageous over all previous studies [38]–[42] on properly modeling the relationship between break type frequencies and *SR*. Therefore, we believe that the use of the SR-dependent CART in the current study is novel and promising. The break type of each syllable juncture is firstly predicted using the refined break-syntax model and the given speaking rate $x = 1/SR$ by

$$B_n^* = \arg\max_{B_n} P_{refined}(B_n | L_n, x) \tag{21}$$

where $P_{refined}(B_n|L_n,x)$ is the refined break-syntax model obtained by further growing of the break-syntax model $P(B_n|L_n,x)$ obtained in Section III. This refinement is motivated by the need of a sophisticated decision tree to more accurately predict break types from linguistic features in the test phase without the help of prosodic-acoustic features. The refinement is realized via applying the same CART algorithm described in Sub-section III-C to the existing tree of the original model $P(B_n|L_n,x)$ with a relaxed split criterion of a smaller ML gain and a smaller minimum sample size so as to let the tree grow deeper.

After performing the break prediction, we then execute the prosodic state prediction by using the three prosodic state sub-models and the three newly-trained prosodic state-syntax models $p(p_n|L_n)$, $p(q_n|L_n)$ and $p(r_n|L_n)$:

$$\begin{aligned}\mathbf{p}^*,\mathbf{q}^*,\mathbf{r}^* = &\arg\max_{\mathbf{p},\mathbf{q},\mathbf{r}} P(p_1|bin(x))P(q_1|bin(x))P(r_1)\\ &\cdot\prod_{n=2}^{N}\begin{pmatrix}P(p_n|p_{n-1},B^*_{n-1}, & bin(x))\\ \cdot P(q_n|q_{n-1},B^*_{n-1}, & bin(x))\cdot P(r_n|r_{n-1},B^*_{n-1})\end{pmatrix}\\ &\cdot\prod_{n=1}^{N}(P(p_n|L_n)P(q_n|L_n)P(r_n|L_n))\end{aligned} \tag{22}$$

where $bin(x)$ is a bin of the histogram of $x(k)$ defined in Eq. (19). Here, the three prosodic state-syntax models are incorporated into the prosodic state prediction because they additionally introduce linguistic information to assist in predicting the three types of prosodic-state tags of pitch, duration and energy level. In this study, the three new models are respectively obtained by the CART algorithm [24] using the prosodic-state labeling results of the training data set. The node splitting criterion is also the ML gain with a minimum sample size constraint.

After the predictions of break type and prosodic state, the SR-dependent prosodic-acoustic feature generation is performed in two steps: (1) reconstruction of normalized prosodic-acoustic features, and (2) SR-denormalization of prosodic-acoustic features. First, the normalized versions of syllable pitch contour, syllable duration, syllable energy level, and syllable-juncture pause duration are generated by

$$\mathbf{sp}^*_n = \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p^*_n} + \boldsymbol{\beta}^f_{B^*_{n-1},t^n_{n-1}} + \boldsymbol{\beta}^b_{B^*_n,t^{n+1}_n} + \boldsymbol{\mu}_{sp'} \tag{23}$$

$$sd^*_n = \gamma_{t_n} + \gamma_{s_n} + \gamma_{q^*_n} + \mu_{sd'} \tag{24}$$

$$se^*_n = \omega_{t_n} + \omega_{f_n} + \omega_{r^*_n} + \mu_{se'} \tag{25}$$

$$pd^*_n \equiv \mu^*_n = \alpha^*_n\beta^*_n \tag{26}$$

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 much degradation. The pause duration $pd^*_n$ is generated from the mean of the Gamma distribution with parameters $\alpha^*_n$ and $\beta^*_n$ found from the leaf node of the break-acoustic model specified by the predicted break type $B^*_n$ and the contextual features $L_n$.

The final four prosodic-acoustic features are then 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, i.e.,

$$\widehat{\mathbf{sp}}_n(i) = \frac{\mathbf{sp}^*_n(i) - \mu^{sp}_g(t_n,i)}{\sigma^{sp}_g(t_n,i)}\tilde{\sigma}^{sp}(x,t_n,i) + \tilde{\mu}^{sp}(x,t_n,i) \quad \text{for } i = 1,2,3,4 \tag{27}$$

$$\widehat{sd}_n = \left(sd^*_n - \mu^{sd}_g\right)/\sigma^{sd}_g \cdot \tilde{\sigma}^{sd}(x) + \mu^{sd}(x) \tag{28}$$

$$\widehat{se}_n = se^*_n \tag{29}$$

$$\widehat{pd}_n = G^{-1}(G(pd^*_n;\alpha^{pd}_g,\beta^{pd}_g);\tilde{\alpha}^{pd}(x),\tilde{\beta}^{pd}(x)) \tag{30}$$

It is noted that $\mu^{sd}(x) = x$ and the reconstructed energy level is not SR-dependent. By incorporating these prosodic feature estimates with the spectral parameters generated from the HMM-based speech synthesizer [31], the synthetic speech can be generated.

*2) Speech Synthesis:* An HMM-based synthesizer is constructed, by using the HTS-2.2 toolkit [34] with the training set of the normal-rate speech corpus (containing 52,192 syllables), for providing the spectral features and the state duration information to the proposed SR-controlled Mandarin TTS system. Many literatures in Mandarin speech synthesis have reported that sub-syllable units of *initials* and *finals* were taken as basic synthesis units of Mandarin HTS systems [43][44]. Those literatures generally agreed that an HMM with five states, left-to-right transition and diagonal covariance matrix was suitable for modeling phonetic variation of an *initial* or a *final*. Therefore, in this study, we also adopt this HMM topology to model sub-syllable synthesis units of 21 *initials* and 39 *finals*. Speech signal is converted into a sequence of 25-dimensional mel-generalized cepstral (MGC) (including the 0th coefficient) vectors in 5 ms interval. The context-dependent HMM (CD-HMM) training for speech synthesis [32] is adopted to simultaneously construct spectral, F0 and state duration models. Note that the F0 model and the state duration model trained by the HTS toolkit are respectively used in our system to determine the voiced/unvoiced status of an HMM state and the state durations of a syllable with duration being predicted by our method. A question set containing 399 questions for decision tree-based context clustering of HMMs is formed from the following features: (1) left, current and right *initial/final* types, (2) contextual break types, and (3) prosodic state. After training, 511, 2,063 and 9,672 leaf nodes are obtained for the decision trees of state duration, MGC and F0, respectively.

The synthesized speech is generated by a modified version of the HTS engine API of the HTS-2.2 toolkit [34]. Several modifications of the original HTS engine API are performed to make it operable on the syllable-based prosodic-acoustic features provided by the proposed prosody generation method. First, durations of HMM states in a syllable or an inter-syllable pause are determined by maximizing the summed log likelihood of Gaussian distributions in the state duration model of the HTS under the constraint that the sum of state durations in a syllable/pause duration equals the predicted syllable duration $\widehat{sd}_n$ or the predicted pause duration $\widehat{pd}_n$. Second, the multi-space distribution (MSD) F0 model in the HTS provided the information about the voiced/unvoiced indicators of the HMM states. Hence, the length and place of syllable pitch contour can be simply determined using the information of the estimated state

TABLE III
CONFUSION MATRIX OF BREAK TYPE PREDICTION FOR THE TEST SET. (UNIT:%)

| TAR\PRE | *B*0 | *B*1 | B2-1 | B2-2 | *B*2-3 | *B*3 | *B*4 | 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** | | | | | | | | |

TABLE IV
RMSES OF FOUR PROSODIC-ACOUSTIC FEATURES ESTIMATED BY SR-HPM AND HTS

| | SR-HPM | | HTS |
|---|---|---|---|
| | Using predicted break | Using correct break | |
| Syllable duration (*ms*) | 48.9 | 48.2 | 55.3 |
| Syllable pitch contour (*log-Hz*) | 0.18 | 0.17 | 0.33 |
| Syllable energy level (*dB*) | 3.64 | 3.55 | - |
| Juncture pause duration (*ms*) | 88.5 | 55.0 | 284.2 |

durations and the voiced/unvoiced indicators. Using the generated syllable pitch contour parameter $\widehat{\mathbf{sp}}_n$, the frame-based pitch contour of syllable $n$ can be reconstructed by orthogonal expansion [22]. Third, the maximum energy calculated from the spectral features (i.e., MGC) in a syllable CD-HMM (i.e., an *initial* CD-HMM connecting with a *final* CD-HMM) is scaled to the predicted energy level $\widehat{se}_n$ before executing the parameter generation algorithm [32] so as to make the generated energy contour smooth and approximate the desired syllable energy levels.

### C. Experimental Results

*1) Objective Evaluations:* The speaking rate-controlled Mandarin TTS system was implemented by the training processes discussed in Subsection IV-B using the SR-specific feature normalization functions, the SR-HPM and the prosody labeling results of the training set obtained in Subsection III-D. We note that the labeling of the training set was done automatically by the modified PLM algorithm. The training set was firstly used to generate the refined break-syntax model to make its leaf nodes increase from 42 to 265. The questions used in the CART algorithm [24] for tree growing were contextual linguistic features related to POS, PM, word length, and syllable's *initial* type. The total number of questions was 336. Table III shows the confusion matrix of the break type prediction for the test set. It can be seen from the table that the predictions for $B0$, $B1$ and $B4$ were good, while all others were fair or poor. The overall accuracy rate was 71.1%. By more detailed analyses, we find that $B3$ was mainly misclassified as $B4$ and $B2$-2. Due to the similarities of acoustic characteristics and functionalities for $B3$ and $B4$, and for $B3$ and $B2$-2, these prediction errors were not fatal. Other serious errors were to classify $B2$-2 as $B3$ and $B2$-1; $B2$-1 as $B1$ and $B2$-2; and $B2$-3 as $B1$, $B2$-2 and $B2$-1. These prediction errors were also not fatal. The real effect of the break type prediction on the SR-controlled TTS will be further evaluated latter via checking the accuracy of generating the four prosodic-acoustic features of syllable pitch contour, syllable duration, syllable energy level and syllable-juncture pause duration.

We then trained the three prosodic state-syntax models, $p(p_n|L_n)$, $p(q_n|L_n)$ and $p(r_n|L_n)$, by the CART algorithm. The question set contained in total 536 questions formed by contextual linguistic features related to POS, PM, word length, and syllable's *initial* type. The numbers of leaf nodes for the resulting $p(p_n|L_n)$, $p(q_n|L_n)$ and $p(r_n|L_n)$ were 273, 244 and 269, respectively. Using these three prosodic state-syntax models and the prosodic state model of SR-HPM, we can predict the three types of prosodic states for each syllable by Eq. (22).

Lastly, the four prosodic-acoustic features were firstly constructed by Eqs. (23)–(26) using the predicted break types and prosodic states as well as the linguistic features specified in these 4 equations, and then denormalized by Eqs. (27)–(30) to generate the final values. Table IV displays the performances of prosodic-acoustic feature predictions for the test set. RMSEs of 48.9 ms, 0.18 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 fatal break type prediction errors, say between non-pause breaks of ($B0$, $B1$, $B2$-1, $B2$-3) and long-pause breaks of ($B3$, $B4$), were found from Table III. By a more detailed analysis, we found that the large pause duration error mainly resulted from the pair-wise confusions of ($B3$, $B4$) and ($B2$-2, $B3$). Actually, these two types of confusions were not perceptually annoying.

For performance comparison, an HMM-based Speech Synthesis System (HTS) [31]–[33] was implemented by the HTS-2.2 toolkit [34] using the normal speech corpus. The same 60 sub-syllables, including 21 *initials* and 39 *finals*, were also taken as basic synthesis units and modeled by five-state, left-to-right transition HMMs. The question set with 1,179 questions for decision tree-based context clustering of HMMs was formed from the contextual linguistic features related to POS, PM, lexical word (LW) length, lexical tone, and syllable's *initial*/*final* type. The trained decision trees for state duration, MGC and F0 contained respectively 774, 2,835, and 10,171 leaf nodes. The synthesized speech utterances of the baseline HTS were generated by the HTS engine API in the HTS-2.2 toolkit [34]. Via setting various $\rho$ factors, speaking rates of the synthesized speech can be controlled.

The results of prosodic-acoustic feature estimation by the HTS system are also listed in Table IV. It can be found from the table that the SR-HPM model outperformed the HTS system on the predictions of syllable duration, syllable pitch contour, and syllable-juncture pause duration. It is noted that the RMSE of energy of the HTS system was not calculated because of the use of global variance.

A typical example of the estimated syllable pitch level and syllable duration for four speaking rates of fast ($SR = 5.6$ syl/sec), normal ($SR = 5.0$ syl/sec), medium ($SR = 4.2$ syl/sec), and slow ($SR = 3.7$ syl/sec) is shown in Fig. 13. It can be found from these two figures that most

TABLE V
THE RESULTS OF MOS TEST

| *SR* (*syl/sec*) | | 6.7 | 5.9 | 5.3 | 4.8 | 4.3 | 4.0 | 3.6 | 3.3 |
|---|---|---|---|---|---|---|---|---|---|
| MOS | PROPOSED | 3.62 | 3.60 | 3.58 | 3.62 | 3.62 | 3.73 | 3.83 | 3.79 |
| | HTS | 3.21 | 3.35 | 3.35 | 3.43 | 3.29 | 3.17 | 3.09 | 3.09 |
| *P*-VALUE | | 0.0018 | 0.037 | 0.027 | 0.049 | 0.0032 | 0.0032 | <0.001 | <0.001 |

TABLE VI
THE RESULTS OF PREFERENCE TEST

| *SR* (*syl/sec*) | | 6.7 | 5.9 | 5.3 | 4.8 | 4.3 | 4.0 | 3.6 | 3.3 |
|---|---|---|---|---|---|---|---|---|---|
| PREFER (%) | PROPOSED | 58.7 | 54.2 | 52.9 | 49.8 | 58.2 | 71.5 | 79.5 | 79.6 |
| | HTS | 20 | 29.3 | 29.3 | 30.7 | 26.7 | 16.9 | 9.8 | 11.1 |
| | EQUAL | 21.3 | 16.5 | 17.8 | 19.5 | 15.1 | 11.6 | 10.7 | 9.3 |
| *P*-VALUE | | <0.001 | 0.006 | 0.002 | 0.009 | <0.001 | <0.001 | <0.001 | <0.001 |

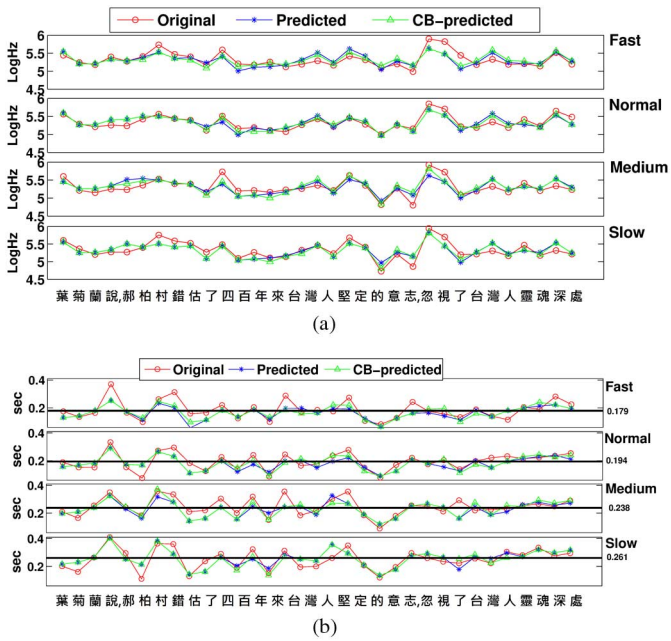



Fig. 13. An example of estimated (a) syllable pitch level and (b) syllable duration for four speaking rates of fast ($SR = 5.6$ syl/sec), normal ($SR = 5.0$ syl/sec), medium ($SR = 4.2$ syl/sec) and slow ($SR = 3.8$ syl/sec). "CB-predicted" denotes prediction using correct break tags. The four horizontal lines in (b) are the syllable duration means of original utterances.

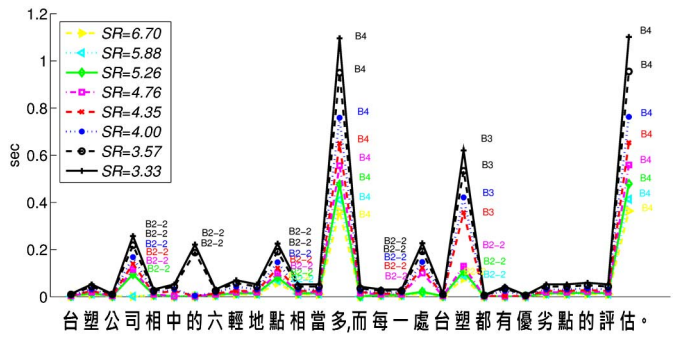



Fig. 14. An example of the break type predictions and their pause duration generations for parts of 8 synthesized utterances.

estimated values matched well with their original counterparts. So, the predictions of these two features were reasonably good.

An example of break type prediction for a paragraph with 8 different speaking rates is given in Fig. 14. The figure 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- and long-pause breaks (i.e., $B2$-2, $B3$, $B4$) were found as *SR* decreased, but also their pause durations increased as *SR* decreased. These results matched with the prior knowledge about the relationship between syllable juncture break pause and speaking rate [1], [26]. To the best of our knowledge, this sophisticated pause generation is a distinct feature of the proposed system not found in all other existing systems.

*2) Subjective Evaluations:* Two subjective tests were conducted to examine the naturalness of the synthesized speech. One was Mean Opinion Score (MOS) in which listeners rated each utterance on a scale from 1 (bad) to 5 (excellent). Another was Preference Test in which two synthetic utterances of a text, generated by the proposed system and the HTS system, were rated with 3 scores representing "prefer A", "prefer B" and "equal". 15 subjects were involved in these two tests. They were all graduate students. 15 short paragraphs with length from 24 to 45 syllables were selected from the outside test data set.[1] These two subjective tests were performed simultaneously in which each subject was asked to give MOS and preference scores to the two utterances of each paragraph synthesized by the proposed system and the HTS system. In each test, the two utterances were randomly assigned as A and B. The original speech of the normal speaking rate was always provided to the subject for his reference. Table V lists the average MOS scores of the two methods for eight speaking rates varying from very fast ($SR = 6.7$ syl/sec) to very slow ($SR = 3.3$ syl/sec). As shown in the table, the MOS scores of the proposed system were in the range of 3.58-3.83, while they were in 3.09-3.43 for the HTS system. The $t$-test was used to measure the significance of the difference between the MOS scores of the two systems. The difference was significant ($< 0.05$) for the three cases of $SR = 5.9$, 5.3 and 4.8 syl/sec; highly significant ($< 0.01$) for $SR = 6.7$, 4.3 and 4.0 syl/sec; and extremely significant ($< 0.001$) for $SR = 3.6$ and 3.3 syl/sec. Table VI lists the results of the preference test. It shows that the proposed system had higher

[1]This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes the synthesized speeches of one test short paragraph, totally 16 WAV-format sound clips, which show the synthesized speech samples of both proposed and baseline HTS systems for 8 different speaking rates. This material is 14.46 MB in size.

preference scores (49.8%-79.6%) than those (9.8%-30.7%) of the HTS system with highly or extremely significant. These results confirmed the effectiveness of the speaking rate control method of the proposed system.

## V. Conclusions

A new approach to modeling the influences of speaking rate on Mandarin speech prosody has been discussed. It provided a systematic way to automatically construct a speaking rate-dependent hierarchical prosodic model (SR-HPM) from a large speech database containing utterances of various speaking rates without human prosody labeling. Experimental results confirmed that the SR-HPM interpreted well the effects of speaking rate on many prosodic phenomena of Mandarin speech. A speaking rate-controlled Mandarin TTS system designed based on the SR-HPM has been realized to illustrate the effectiveness of the speaking rate modeling. 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. By two subjective tests, the system was shown to outperform the popular HTS method significantly. High performance of the speaking rate control method of the system has therefore been confirmed.

## Acknowledgment

The authors would like to thank the ACLCLP for providing the Treebank Corpus.

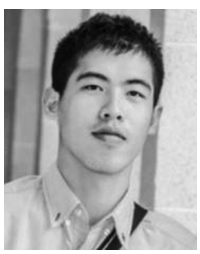
## References

[1] A. Li and Y. Zu, "Speaking rate effects on discourse prosody in standard Chinese," in *Proc. Speech Prosody*, May 2008, pp. 449–452.
[2] S. A. Jun, "The effect of phrase length and speech rate on prosodic phrasing," in *Proc. ICPhS*, Barcelona, Spain, 2003, pp. 483–486.
[3] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proc. ICSLP*, Oct. 2000, pp. 362–365.
[4] S. M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *Proc. ICASSP*, 2010, pp. 4306–4309.
[5] H. Fujimura, T. Masuko, and M. Tachimori, "A duration modeling technique with incremental speech rate normalization," in *Proc. INTERSPEECH'10*, Sep. 2010, pp. 2962–2965.
[6] D. Jouvet, D. Fohr, and I. Illina, "About handling boundary uncertainty in a speaking rate dependent modeling approach," in *Proc. INTERSPEECH'11*, Aug. 2011, pp. 2593–2596.
[7] T. Shinozaki and S. Furui, "Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation," in *Proc. ASRU'03*, Nov. 2003, pp. 417–422.
[8] J. Zheng, H. Franco, and A. Stolcke, "Rate-of-speech modeling for large vocabulary conversational speech recognition," in *Proc. ASRU'00*, Sep. 2002, pp. 145–149.
[9] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 391–400, Jul. 2004.
[10] C. Y. Chiang, C. C. Tang, H. M. Yu, Y. R. Wang, and S. H. Chen, "An investigation on the mandarin prosody of a parallel multi-speaking rate speech corpus," in *Proc. Oriental COCOSDA'09*, Aug. 2009, pp. 148–153.
[11] 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'11*, Aug. 2011, pp. 1845–1848.
[12] Y. Zu, A. Li, and Y. Li, "Speech rate effects on prosodic features," Report of Phonetic Research 2006 Inst. of Linguist., Chinese Acad. Soc. Sci., pp. 141–144.
[13] C. H. Hsieh, C. Y. Chiang, Y. R. Wang, H. M. Yu, and S. H. Chen, "A new approach of speaking rate modeling for mandarin speech prosody," in *Proc. INTERSPEECH'12*, Portland, OR, USA, Aug. 2012, Tue.P3a.03.
[14] S. H. Chen, C. H. Hsieh, C. Y. Chiang, H. C. Hsiao, Y. R. Wang, and Y. F. Liao, "A speaking rate-controlled mandarin TTS system," in *Proc. ICASSP'13*, Vancouver, BC, Canada, May 2013, pp. 6900–6903.
[15] K. Iwano, M. Yamada, T. Togawa, and S. Furui, "Speech-rate variable HMM-based Japanese TTS system," in *Proc. TTS'02*, Sep. 2002.
[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," *ICSLP '98*, pp. 29–32, 1998.
[17] K. U. Ogbureke, J. P. Cabral, and J. Carson-Berndsen, "Explicit duration modeling in HMM-based speech synthesis using a hybrid hidden Markov model-multilayer perceptron," in *Workshops Statist. Percept. Audition Speech Commun. Adaptive Learn. (SCALE)*, 2012.
[18] 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. EMBC'06*, Sep. 2006, pp. 5691–5694.
[19] 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'10*, Sep. 2010, pp. 2186–2189.
[20] R. Srikanth, N. Bajibabu, and K. Prahallad, "Duration modeling in voice conversion using artificial neural networks," in *Proc. 19th Int. Conf. Syst., Signals, Image Process. (IWSSIP)*, 2012, pp. 556–559.
[21] 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. Amer.*, vol. 125, no. 2, pp. 1164–1183, Feb. 2009.
[22] 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, Sep. 1990.
[23] 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.
[24] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Tree*. Belmont, CA, USA: Wadsworth, 1984.
[25] "Sinica Treebank 3.0," [Online]. Available: http://www.aclclp.org.tw/use_stb.php
[26] C.-Y. Tseng, "Corpus phonetic investigations of discourse prosody and I-ligher level information," (in Chinese) *Lang. Linguist.*, vol. 9, no. 3, pp. 659–719, 2008.
[27] C. C. Hsia, C. H. Wu, and J. Y. Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1994–2003, Aug. 2010.
[28] Y. Qian, Z. H. Wu, B. Y. Gao, and F. K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1702–1710, Aug. 2011.
[29] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 6, no. 3, pp. 226–269, May 1998.
[30] S. H. Chen, W. H. Lai, and Y. R. Wang, "A new duration modeling approach for Mandarin speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 308–320, Jul. 2003.
[31] 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.
[32] T. Yoshimura, "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems," Ph.D. dissertation, Nagoya Inst. of Technol., Nagoya, Japan, Jan. 2002.
[33] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, Bonn, Germany, Aug. 2007, pp. 294–299.
[34] The HTS working group, "HTS-2.2 source code and demonstrations," [Online]. Available: http://hts.sp.nitech.ac.jp/?Download
[35] K. Hirose, H. Lei, and H. Fujisaki, "Analysis and formulation of prosodic features of speech in standard Chinese based on a model of generating fundamental frequency contours," *J. Acoust. Soc. Jpn.*, vol. 50, no. 3, pp. 177–187, 1994.
[36] G. P. Chen, G. Bailly, Q. F. Liu, and R. H. Wang, "A superposed prosodic model for Chinese text-to-speech synthesis," in *Proc. ISCSLP'04*, Dec. 2004, pp. 117–120.
[37] G. Bailly and B. Holm, "SFC: A trainable prosodic model," *Speech Commun.*, vol. 46, no. 3-4, pp. 348–364, Jul. 2005.
[38] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," *Comput. Linguist.*, vol. 20, pp. 27–52, 1994.
[39] H. J. Peng, C. C. Chen, C. Y. Tseng, and K. J. Chen, "Predicting prosodic words from lexical words-A first step towards predicting prosody from text," in *ISCSLP'04*, 2004, pp. 173–176.

[40] M. Chu and Y. Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," *Computat. Linguist. and Chinese Lang. Process.*, vol. 6, pp. 61–82, 2001.
[41] D. W. Xu, H. F. Wang, G. H. Li, and T. Kagoshima, "Parsing hierarchical prosodic structure for Mandarin speech synthesis," in *Proc. ICASSP*, 2006, vol. 1, pp. 14–19.
[42] A. W. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," in *Proc. Eurospeech*, 1997, pp. 995–998.
[43] Q. S. Duan, S. Y. Kang, Z.-Y. Wu, L. H. Cai, Z. W. Shuang, and Y. Qin, "Comparison of syllable/phone HMM based mandarin TTS," in *Proc. ICPR'10*, Aug. 2010, pp. 4496–4499.
[44] Y. F. Liao, S. H. Lyu, and M. L. Wu, "The NTUT Blizzard Challenge 2010 Entry," in *Blizzard Challenge 2010 Workshop*, 2010 [Online]. Available: http://festvox.org/blizzard/blizzard2010.html

**Sin-Horng Chen** (SM'94) received the B.S. degree in communication engineering and the M.S. degree in electronics engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, Lubbock, in 1983.

He became an Associate Professor and a Professor in the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively. His major research interest is in speech signal processing, especially in Mandarin speech recognition and text-to-speech.



**Chiao-Hua Hsieh** received the B.S. degree in electrical engineering from Taipei University of Technology, Taipei, Taiwan, in 2010, and the M.S. degree in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2012.

He is an Engineer with the Media Tech Company, Hsinchu, Taiwan. His major research area is prosody assisted speech synthesis.

**Chen-Yu Chiang** (M'09) was born in Taipei, Taiwan, in 1980. He received the B.S., M.S., Ph.D. degrees in communication engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2002, 2004 and 2009, respectively.

In 2009, he was a Postdoctoral Fellow at the Department of Electrical Engineering, NCTU, where he primarily worked on prosody modeling for automatic speech recognition and text-to-speech system, under the guidance of Prof. Sin-Horng Chen. He was also a Visiting Scholar at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta. His main research interests are in speech processing, in particular prosody modeling, automatic speech recognition and text-to-speech systems.

Dr. Chen-Yu Chiang is a member of ISCA and ASA.I.

**Hsi-Chun Hsiao** received the B.S. degree in electrical engineering from National Center University, Zhongli, Taiwan, in 2003, and the M.S. degree in communication engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2005. He was a Ph.D. candidate at National Chiao Tung University in 2007. He is currently a researcher with the Value Creation Center, Wistron Co., Taiwan. His major research areas are speech signal processing and image processing.

**Yih-Ru Wang** (M'06) received the B.S. and M.S. degree from the Department of Communication Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, NCTU, in 1995.

He was an Instructor in the Department of Communication Engineering, NCTU, from 1987 to 1995. In 1995, he became an Associate Professor. His general research interests are automatic speech recognition and computational linguistics.

**Yuan-Fu Liao** received the B.S., M.S., and Ph.D. degrees from National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1991, 1993, and 1998, respectively. From January 1999 to June 1999, he was a Postdoctoral Researcher with the Department of Communication Engineering, National Chiao-Tung University. From September 1999 to February 2002, he became a Research Engineer with Philips Research East Asia, Taiwan, Since February 2002, he has been with the Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan, where he is currently an associate Professor. His major research interest is in speech signal processing, especially, speech recognition and speech synthesis.

Prof. Yuan-Fu Liao is a member of IEEE, ISCA and ACLCLP.

**Hsiu-Min Yu** received her M.A. in Linguistics from the Institute of Linguistics, Fu-Jen Catholic University, Taipei, Taiwan, in 1984. In 1984 she worked as an English lecturer, teaching Freshman English for non-English majors at Tatung Institute of Technology. From 1985 to 1990, she was a full-time assistant researcher at Telecommunication Laboratories, Ministry of Transportation and Communication, doing researches on Mandarin text-to-speech and speech-recognition technology. Since 1990, she has been teaching English courses in Chung Hua University as a full-time lecturer, and is now also pursuing the Ph.D. program offered by the Graduate Institute of Taiwan Languages and Language Education, National Hsinchu University of Education.