STRUCTURAL MAXIMUM A POSTERIORI SPEAKER ADAPTATION OF SPEAKING RATE-DEPENDENT HIERARCHICAL PROSODIC MODEL FOR MANDARIN TTS

I-Bin Liao[1, 3], *Chen-Yu Chiang*[2], *Sin-Horng Chen*[3]

[1]TL of Chunghwa Telecom Co., Ltd
[2]Dept. of Communication Engineering, National Taipei University, Taiwan
[3]ECE Department of NCTU

snet@cht.com.tw,cychiang@mail.ntpu.edu.tw,schen@mail.nctu.edu.tw

Abstract

In this paper, a structural maximum a posterior speaker adaptation method to adjust the existing speaking rate (SR) dependent hierarchical prosodic model (SR-HPM) to a new speaker's data for realizing a new voice of any given SR is discussed. The adaptive SR-HPM is formulated based on MAP estimation with a reference SR-HPM serving as an informative prior. The prior information provided by the reference SR-HPM is hierarchically organized by decision trees. The results of objective and subjective evaluations showed that the proposed method not only performed slightly better than the maximum likelihood-based model in the observed SR range of the target speaker's data, but also was much better in the unseen SR range.

Index Terms— speaker adaptation, hierarchical prosodic model, prosodic-acoustic features, Mandarin TTS

1. Introduction

In the past, we have developed a speaking rate (SR)-controlled Mandarin TTS system (MTTS) using a large speech corpus containing utterances of various SRs of a female speaker. An SR-HPM [4] was trained and used in the TTS system to generate prosodic-acoustic features for any given SR [3]. A previous study [1] was then conducted to adapt the SR-HPM to a small dataset of a new speaker for generating a new voice. The effectiveness of four feature Normalization Functions (NFs) on generating prosodic-acoustic features outside the observed SR range of the target speaker's data was studied. Two issues were raised:1) sparseness of adaptation data due to a large space of the model parameters, and 2) poor estimation of prior variances due to the fact that only one Mandarin SR-HPM is trained from the speech corpus of one speaker.

Many speaker adaptation techniques have been proposed in the past. Among them, maximum likelihood linear regression (MLLR) [5] and maximum a posteriori (MAP) [6] are two popular approaches for spectral model adaptation. In this study, a structural MAP (SMAP) adaptation method [2] is proposed to tackle these two issues. The speaker adaptation for the SR-HPM works in a similar flow as the training of the original SR-HPM but in an adaptive fashion.

The remainder of the paper is organized as follows. Section 2 gives a brief review of the existing SR-controlled MTTS. Section 3 presents the proposed SMAP speaker adaptation method. Experimental results are discussed in Section 4. Some conclusions and future works are given in the last section.

2. A Review of the Existing SR-Controlled MTTS

Fig.1 displays a block diagram of the existing SR-controlled Mandarin TTS system [4,14]. The system can be divided into three parts: 1) text analysis (TA), 2) prosody generation (PG), and 3) speech synthesis. The TA produces linguistic feature, **L**, from raw text. The PG generates the four prosodic-acoustic features of syllable log-F0 contour (**sp**), syllable duration (**sd**), syllable energy level (**se**), and inter-syllable pause duration (**pd**) based on a trained SR-HPM with a specified SR, i.e. *x*. Last, speech is synthesized by an HMM-based speech synthesizer given with the generated prosodic-acoustic features.

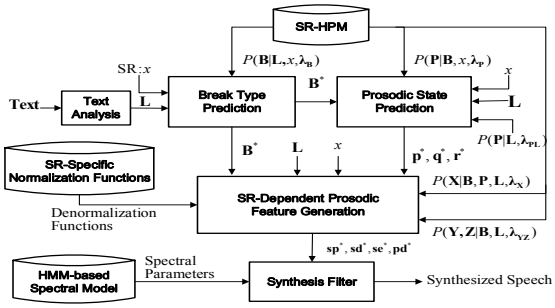

Figure 1: A block diagram of the existing SR-controlled Mandarin TTS system.

2.1. SR-controlled prosody generation

In the PG, the prosodic structure in terms of a break type sequence is first generated by the SR-dependent break-syntax model, $P(\mathbf{B}|\mathbf{L},x,\boldsymbol{\lambda}_{\mathbf{B}})$, which is implemented by a modified CART decision tree (DT) [13] in which probability of each break type is a linear function of SR, *x*:

$$B_n^* = \arg\max_{B_n} P(B_n|L_n,x,\boldsymbol{\lambda}_{\mathbf{B}}) \quad (1)$$

where *n* represent syllable index; B_n is the break type of inter-syllable juncture following syllable *n*; L_n is the contextual linguistic features of syllable *n*; $\boldsymbol{\lambda}_{\mathbf{B}}$ is the model parameters of the DT of the break-syntax model. Then, a prosodic state sequence, $\mathbf{P}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$, is generated by the prosodic state model, $P(\mathbf{P}|\mathbf{B},x,\boldsymbol{\lambda}_{\mathbf{P}})$, and the prosodic state-syntax model, $P(\mathbf{P}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{PL}})$, with linguistic features **L** and SR:

$$\mathbf{p}^*,\mathbf{q}^*,\mathbf{r}^* = \arg\max_{\mathbf{p},\mathbf{q},\mathbf{r}} \begin{bmatrix} P(\mathbf{p}|\mathbf{B}^*,x,\boldsymbol{\lambda}_{\mathbf{p}})P(\mathbf{q}|\mathbf{B}^*,x,\boldsymbol{\lambda}_{\mathbf{q}})P(\mathbf{r}|\mathbf{B}^*,x,\boldsymbol{\lambda}_{\mathbf{r}}) \\ P(\mathbf{p}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{pL}})P(\mathbf{q}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{qL}})P(\mathbf{r}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{rL}}) \end{bmatrix} \quad (2)$$

where $P(u|\mathbf{B}^*,x,\boldsymbol{\lambda}_u)|u=\mathbf{p},\mathbf{q},\mathbf{r}$ are three sub-models of $P(\mathbf{P}|\mathbf{B},x,\boldsymbol{\lambda}_{\mathbf{P}})$ for pitch (**p**), duration (**q**), and energy (**r**) prosodic states, respectively; $P(u|\mathbf{L},\boldsymbol{\lambda}_{uL})|u=\mathbf{p},\mathbf{q},\mathbf{r}$ are three sub-models of $P(\mathbf{P}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{PL}})$. After obtaining the prosodic structure in terms of $\mathbf{B}^*,\mathbf{p}^*$, $\mathbf{q}^*$, and $\mathbf{r}^*$, the four SR-normalized prosodic-acoustic features can be generated by the syllable prosodic-acoustic model, $P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})$, and the syllable juncture prosodic-acoustic model, $P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{YZ}})$:

$$\begin{aligned}
\mathbf{sp}'_n &= \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p^*_n} + \boldsymbol{\beta}^f_{B^*_{n-1},tp_{n-1}} + \boldsymbol{\beta}^b_{B^*_n,tp_n} + \boldsymbol{\mu}_{sp} \\
sd'_n &= \gamma_{t_n} + \gamma_{s_n} + \gamma_{q^*_n} + \mu_{sd} \\
se'_n &= \omega_{t_n} + \omega_{f_n} + \omega_{r^*_n} + \mu_{se} \\
pd'_n &= \kappa_{B^*_k,L_k}\,\theta_{B^*_k,L_k}
\end{aligned} \tag{3}$$

where $\boldsymbol{\beta}'s$, $\gamma's$, and $\omega's$ are affecting patterns (APs) respectively for **sp**, **sd**, **se**, included in the model parameter set of $\boldsymbol{\lambda}_{\mathbf{X}}$. These APs are associated with tone t_n, base-syllable type s_n, final type f_n, prosodic state $\{p_n, q_n, r_n\}$, and forward/backward pitch coarticulations $\boldsymbol{\beta}^f_{B,tp}$ / $\boldsymbol{\beta}^b_{B,tp}$ conditioned on adjacent break type and tone pair $tp_n = (t_n, t_{n+1})$. The parameters $\kappa's$ and $\theta's$ are the parameters of the Gamma distribution for *pd* found from the leaf nodes of the syllable-juncture model given with the predicted break $B^*_{k,n}$ and the linguistic features $L_{n,k}$. They are included in the model set of $\boldsymbol{\lambda}_{\mathbf{YZ}}$. Last, the four prosodic-acoustic features are SR-denormalized by the inverse operation of the NFs:

$$\begin{aligned}
\mathbf{sp}^*_n(i) &= \frac{\mathbf{sp}'_n(i) - \mu^{sp}_g(t_n,i)}{\sigma^{sp}_g(t_n,i)}\tilde{\sigma}^{sp}(x,t_n,i) + \tilde{\mu}^{sp}(x,t_n,i) \\
sd^*_n &= (sd'_n - \mu^{sd}_g)/\sigma^{sd}_g \cdot \tilde{\sigma}^{sd}(x) + \mu^{sd} \\
se^*_n &= se'_n \\
pd^*_n &= G^{-1}(G(pd'_n;\alpha^{pd}_g,\beta^{pd}_g);\tilde{\alpha}^{pd}(x),\tilde{\beta}^{pd}(x))
\end{aligned} \tag{4}$$

where $\mu^{sp}_g(t,i)$ and $\sigma^{sp}_g(t,i)$ are the global mean and standard deviation for the *i*-th dimension of tone *t*; $\tilde{\mu}^{sp}(\cdot)$ and $\tilde{\sigma}^{sp}(\cdot)$ are tone-dependent NFs for the *i*-th log-F0 component; $\{\mu^{sd}_g, \sigma^{sd}_g\}$ and $\{\alpha^{pd}_g, \beta^{pd}_g\}$ are parameters representing the distributions of the SR-normalized syllable duration and pause duration, respectively; $\tilde{\sigma}^{sd}(x)$ and $\{\tilde{\alpha}^{pd}(x), \tilde{\beta}^{pd}(x)\}$ are the NFs for syllable duration and pause duration. It is noted that $\tilde{\mu}^{sp}(\cdot)$ and $\tilde{\sigma}^{sp}(\cdot)$ are two 1st order polynomial functions of SR, *x*, while $\tilde{\sigma}^{sd}(x)$, $\tilde{\alpha}^{pd}(x)$ and $\tilde{\beta}^{pd}(x)$ are three 2nd order polynomial functions of *x*. All the parameters of the NFs are obtained by the mean-square-error (MSE) criterion to fit utterance-wise means and standard deviations of log-F0, syllable duration, or pause duration.

2.2. Training of the SR-HPM

The parameters of SR-HPM are trained by maximizing the following maximum likelihood (ML)-based formula:

$$\boldsymbol{\lambda}^*, \mathbf{T}^* = \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\mathbf{T},\mathbf{A} \mid \mathbf{L},\mathbf{x},\boldsymbol{\lambda}) \tag{5}$$

where $\boldsymbol{\lambda}$ is the model parameters of SR-HPM; $\mathbf{T} = \{\mathbf{B},\mathbf{P}\}$ is the prosody tags of break type **B** and prosodic state **P**; $\mathbf{A} = \{\mathbf{X},\mathbf{Y},\mathbf{Z}\}$ is the set of prosodic-acoustic features: $\mathbf{X} = \{\mathbf{sp},\mathbf{sd},\mathbf{se}\}$, $\mathbf{Y} = \{\mathbf{pd},\mathbf{ed}\}$, and $\mathbf{Z} = \{\mathbf{pj},\mathbf{dl},\mathbf{df}\}$; **ed**, **pj**, **dl** and **df** are energy-dip level, normalized pitch jump, and the two syllable duration lengthening factors; **L** is the linguistic features; and **x** is SR. The term $P(\mathbf{T},\mathbf{A} \mid \mathbf{L},\mathbf{x},\boldsymbol{\lambda})$ is further expressed by the following four models:

$$\begin{aligned}
P(\mathbf{T},\mathbf{A}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}) \approx\ & P(\mathbf{X}|\mathbf{B},\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{YZ}}) \\
& \times P(\mathbf{P}|\mathbf{B},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{PS}})P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}})
\end{aligned} \tag{6}$$

A special designed iterative algorithm, i.e. the PLM algorithm [5] is adopted to simultaneously label the prosody tag sequence **T** and to update the model parameter sets $\boldsymbol{\lambda}$.

3. Adaptation of Normalization Functions

Since the three types of NFs, i.e. $\{\tilde{\mu}^{sp}(\cdot), \tilde{\sigma}^{sp}(\cdot)\}$, $\tilde{\sigma}^{sd}(x)$ and $\{\tilde{\alpha}^{pd}(x), \tilde{\beta}^{pd}(x)\}$ are all modeled by polynomial functions of SR (*x*), the parameters of the each NF could be estimated in the same way, i.e. adaptation by the MAP linear regression (MAPLR) approach. For simplicity, we only illustrate the adaptation for $\tilde{\sigma}^{sd}(x)$ in this section.

For modeling convenience, a variable $x(k)$ representing the average syllable duration of the *k*-th utterance is defined as an independent variable for the NF $\tilde{\sigma}^{sd}(x)$. In the previous study [1], we found that the best result is the smoothed NF passing through the point of the average syllable standard deviation at the average SR of the target speaker corpus. Therefore, the parameters for $\tilde{\sigma}^{sd}(x)$ are obtained by the following MAPLR with a Lagrange multiplier λ :

$$\begin{aligned}
a^*,b^*,c^* &= \arg\max_{a,b,c}\left[\ln P(a,b,c|\boldsymbol{\sigma}^{sd}) + \lambda(\bar{\sigma} - a\bar{x}^2 - b\bar{x} - c)\right] \\
&\approx \arg\max_{a,b,c}\left[\begin{matrix}\ln\left(P(\boldsymbol{\sigma}^{sd}|a,b,c)^{w(\mathbf{x})}P(a,b,c)\right) \\ +\lambda(\bar{\sigma} - a\bar{x}^2 - b\bar{x} - c)\end{matrix}\right]
\end{aligned} \tag{7}$$

where *a*, *b*, and *c* are parameters of the 2nd order polynomial; $P(\boldsymbol{\sigma}^{sd}|a,b,c)$ is the likelihood function modeling the observed target speaker's utterance-wise syllable-duration standard deviations $\boldsymbol{\sigma}^{sd} = \{\sigma^{sd}(k)\}_{k=1\sim K}$; $\sigma^{sd}(k)$ is the observed syllable-duration standard deviation of the *k*-th utterance; $\mathbf{x} = \{x(k)\}_{k=1\sim K}$; and $w(\mathbf{x})$ is a weight to consider the SR coverage of utterances in the whole target speaker corpus; $P(a,b,c) = P(a)P(b)P(c)$ is the prior probability of the polynomial coefficients; $\bar{x}$ and $\bar{\sigma}$ are respectively the average SR and the average syllable-duration standard deviation of the target speaker corpus. The likelihood function is further elaborated by

$$P(\boldsymbol{\sigma}^{sd} \mid a,b,c) = \prod_k N(\sigma^{sd}(k); \tilde{\sigma}^{sd}(x(k)), v^{sd}) . \tag{8}$$

The weight $w(\mathbf{x})$ is defined by

$$w(\mathbf{x}) = std(x(k)) / std(\hat{x}(k)) \tag{9}$$

where $std(x(k))$ and $std(\hat{x}(k))$ are the standard deviations of the observed utterance-wise SR of the target speaker speech corpus and the reference Mandarin speech corpus, respectively. The priors $P(a)$, $P(b)$ and $P(c)$ are all assumed to be Gaussian distributed. The means and variances for the priors are estimated by n-fold sets of the reference speech corpora.

4. Adaptation of SR-HPM

The adaptive SR-HPM is formulated based on the MAP estimation with the reference SR-HPM serving as an informative prior. It is designed to simultaneously estimate the model parameters of target SR-HPM, $\boldsymbol{\lambda}^*$, and label the prosody tags of target speaker, $\mathbf{T}^*$, given with prosodic-acoustic features, $\mathbf{A}$, linguistic features, $\mathbf{L}$, and SR, $\mathbf{x}$:

$$\begin{aligned}
\boldsymbol{\lambda}^*, \mathbf{T}^* &= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\boldsymbol{\lambda} \mid \mathbf{T},\mathbf{A},\mathbf{L},\mathbf{x}) = \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\boldsymbol{\lambda},\mathbf{T},\mathbf{A},\mathbf{L},\mathbf{x}) \\
&= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\mathbf{T},\mathbf{A} \mid \mathbf{L},\mathbf{x},\boldsymbol{\lambda})P(\mathbf{L},\mathbf{x} \mid \boldsymbol{\lambda})P(\boldsymbol{\lambda}) \\
&= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\mathbf{T},\mathbf{A} \mid \mathbf{L},\mathbf{x},\boldsymbol{\lambda})P(\boldsymbol{\lambda})
\end{aligned} \tag{10}$$

where $P(\mathbf{T},\mathbf{A} \mid \mathbf{L},\mathbf{x},\boldsymbol{\lambda})$ is the original SR-HPM in Eq. (6) and $P(\boldsymbol{\lambda})$ is the prior probability for the SR-HPM parameters.

4.1. The proposed adaptive PLM algorithm

The adaptive PLM algorithm is specially designed for training the parameters of SR-HPM in an adaptation fashion. Since the

SR-HPM consists of many sub-models, a sequential optimization procedure is conducted to maximize each part of the model parameters as described as follows:

Step 1: Set all the parameters of SR-HPM as their prior means.

Step 2: Find the optimal break type sequence using the syllable-juncture prosodic-acoustic model and the SR-dependent break-syntax model by

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y,Z}})P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}}) \quad (11)$$

Step 3: Obtain the optimal prosodic state sequence using the syllable prosodic-acoustic and the prosodic state models by

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} P(\mathbf{X}|\mathbf{B}^*,\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\mathbf{P}|\mathbf{B}^*,\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}}) \quad (12)$$

Step 4: Adapt the sets of $\boldsymbol{\lambda}_{\mathbf{X}}$, $\boldsymbol{\lambda}_{\mathbf{Y,Z}}$, $\boldsymbol{\lambda}_{\mathbf{B}}$, and $\boldsymbol{\lambda}_{\mathbf{P}}$ by SMAP:

$$\begin{aligned}
\boldsymbol{\lambda}_{\mathbf{X}}{}^* &= \arg\max_{\lambda_{\mathbf{X}}} P(\mathbf{X}|\mathbf{B}^*,\mathbf{P}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\boldsymbol{\lambda}_{\mathbf{X}}) \\
\boldsymbol{\lambda}_{\mathbf{Y,Z}}{}^* &= \arg\max_{\lambda_{\mathbf{Y,Z}}} P(\mathbf{Y},\mathbf{Z}|\mathbf{B}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y,Z}})P(\boldsymbol{\lambda}_{\mathbf{Y,Z}}) \\
\boldsymbol{\lambda}_{\mathbf{B}}{}^* &= \arg\max_{\lambda_{\mathbf{B}}} P(\mathbf{B}^*|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}})P(\boldsymbol{\lambda}_{\mathbf{B}}) \\
\boldsymbol{\lambda}_{\mathbf{P}}{}^* &= \arg\max_{\lambda_{\mathbf{P}}} P(\mathbf{P}|\mathbf{B}^*,\boldsymbol{\lambda}_{\mathbf{P}})P(\boldsymbol{\lambda}_{\mathbf{P}})
\end{aligned} \quad (13)$$

Step 5: Find the optimal break type sequence using all sub-models of the SR-HPM by

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} \begin{bmatrix} P(\mathbf{X}|\mathbf{B},\mathbf{P}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}}{}^*)P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y,Z}}{}^*) \\ P(\mathbf{P}^*|\mathbf{B},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}}{}^*)P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}}{}^*) \end{bmatrix} \quad (14)$$

Step 6: If a convergence is reached, exit; otherwise set $\boldsymbol{\lambda}_{\mathbf{X}} = \boldsymbol{\lambda}_{\mathbf{X}}{}^*$, $\boldsymbol{\lambda}_{\mathbf{P}} = \boldsymbol{\lambda}_{\mathbf{P}}{}^*$ and go to Step 3.

4.2. Design of the prior probabilities

Two problems were encountered in the MAP estimation of previous study [1]. First of all, due to the large space of the model parameters, the sparse adaptation data would degrade the capability of MAP estimation. Secondly, the variances of the priors are hard to obtain because only one Mandarin SR-HPM is trained from speech corpora of one speaker. To tackle these two problems, the SMAP approach [2] is adopted here to consider using hierarchical structures of model parameter spaces. In this study, hierarchical structures are made by the DT clustering method for model parameter sets of $\{\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}\}$, $\{\gamma_s\}$, $\{\omega_f\}$, $\boldsymbol{\lambda}_{\mathbf{YZ}} = \{\boldsymbol{\lambda}_{pd}, \boldsymbol{\lambda}_{ed}, \boldsymbol{\lambda}_{pj}, \boldsymbol{\lambda}_{dl}, \boldsymbol{\lambda}_{df}\}$ and $\boldsymbol{\lambda}_{\mathbf{B}} = \{c^B_{m,j}, d^B_{m,j}\}$. Note that the parameter sets $\boldsymbol{\lambda}_{\mathbf{YZ}}$ and $\boldsymbol{\lambda}_{\mathbf{B}}$ have been already hierarchically organized by the DT clustering method. We further adopt the CART algorithm to cluster parameter sets of $\{\boldsymbol{\beta}^f_{B,tp}, \boldsymbol{\beta}^b_{B,tp}\}$ and $\{\gamma_s / \omega_f\}$ according to question sets made of tone/break type context properties, and phonetic properties. Therefore, we can directly apply these tree structures to represent the dependencies between model parameters at one level to their adjacent levels in DTs. Note that each node in a DT is modeled by a Gaussian distribution, and served as an informative prior. The means of the priors are directly set to be the parameters of the Mandarin SR-HPM. The variance can be estimated by the means of the priors in the descendant nodes of the current node. The variances in the leaf nodes are directly inherited from their father nodes.

5. Experiments on Adaptions of NFs and SR-HPM

Effectiveness of the proposed method was examined by simulations on the speech corpus of a male target speaker. The adaptation corpus contained 100 paragraphic utterances with 10,559 syllables uttered in a range of moderate SR, i.e. 0.22~0.27 sec/syl. The testing corpus contained 12 utterances with 1163 syllables uttered in a wide range of SR from 0.15 to 0.43 sec/syl. The average length of these utterances was 117 syllables. It is noted that the adaptation corpus is also used to train an HMM-based speech synthesizer by the HTS toolkit [8].

5.1. Examination of adapted NFs

Pause duration (*pd*) and syllable duration (*sd*) are the two most important features when considering SR. Therefore, we select the adapted NFs of *pd* and *sd* for illustration. Fig. 2a plots the NFs of two estimation methods for pause duration. It can be found from the figure that the NF by the MAP estimation (red curve) has the best fit to data in all SR range; while the NF by the MSE estimation (blue curve) has not good fits to data outside the SR range of the training data. Fig. 2b shows that the extrapolation of pause duration for 7 break types looks good to comply with our intuition: pause duration decreases as SR decreases and has larger perceptible values for B3 and B4 in slow speech (say at SR= 0.3 sec/syl).

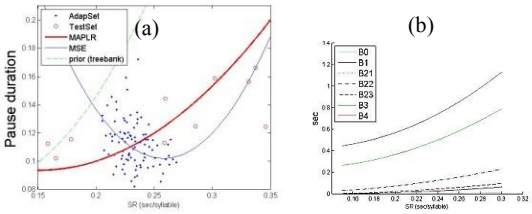

Fig. 2: (a) Pause mean vs. SR, and (b) Average pause durations of 7 break types vs. SR.

Fig. 3 plots the NFs of two estimation methods for syllable duration standard deviation. It can be found from the figure that the NF by the MAP estimation (red curve) fits data well in all SR range; while the NF by the MSE estimation (blue curve) fits data poor outside the SR range of the training data.

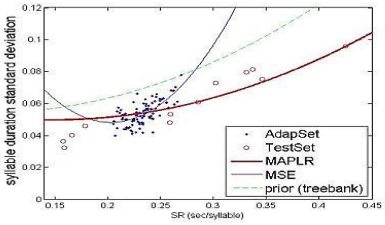

Fig. 3: syllable durations std vs. SR.

5.2. Examination of adapted SR-HPM

Fig. 4 shows the reference and adapted SR-dependent break-syntax models expressed by trees at the average SR (0.215 sec/syl) of the target speaker. Each node lists a histogram of 7 break types in the order of B0, B1, B2-1, B2-2, B2-3, B3, and B4 from left to right. A node in the DTs is split into two son nodes according to the linguistics question listed on it with solid/dotted line corresponding to yes/no to the question. By observing the histograms in the two root nodes, it is found that the two speakers perform different break insertion strategies. Generally, the target speaker inserted fewer B2-2s and B2-3s than the reference speaker did. More interestingly, the target speaker preferred to insert more B2-2s (short pause break) than B3s/B4s (median/long pause break) in word junctures with punctuation marks (PMs). In addition, the target speaker inserted more B3s and B4s than the reference speaker in the non-PM interword case (node 6). The break labeling result of the target speaker is quite reasonable since the target speaker is not a professional announcer and he could not read as fluent

as a professional announcer did. From the above analysis, the proposed adaptive PLM algorithm could learn the distinct break strategy of the target speaker. We confirm this observation by an informal listening test to find that the utterances of the target speaker sound disfluent and not well-planned.

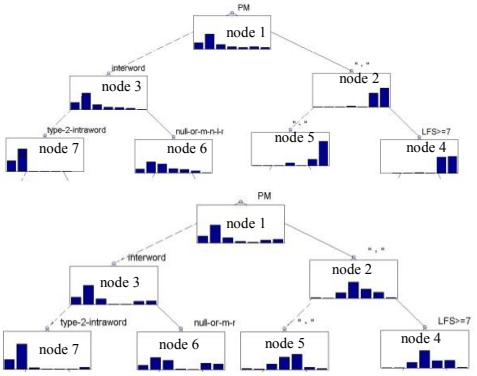

Fig. 4: break syntax tree of reference (upper) and the target speaker (lower)

Table 1 lists modeling errors for different aggregation of APs for **sp**, *sd*, and *se*. The variances of modeling residuals of the three syllable-based prosodic-acoustic features became very small as more APs were considered. Table 2 shows the root mean squared errors (RMSEs) of the four prosodic-acoustic features for three different adaptation data sizes. It is found that RMSEs gradually decreased as the size of data increased for **sp**, *sd*, *pd* and *se*. Generally, the results shown in Table 2 confirm the effectiveness of the proposed approaches of NF adaptation and adaptive SR-HPM for the new speaker. Table 3 shows 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 that RMSEs were large only for B3 and B4 for the proposed SMAP method. Since the pause durations of these break types were inherently longer with much larger dynamic ranges, these results were reasonably good. The overall performance of the proposed methods is better than the ML method.

Table 1: Modeling Errors for different Aggregation of APs. IF:base-syllable type. F:final type. PS:prosodic state. Coar: forward and backward coarticulation. Tone: lexical tone

Log-F0 (sp)		Duration (sd)		Energy (se)	
APs	TRE	APs	TRE	APs	TRE
+Tone	56.1%	+Tone	91.7%	+Tone	84.8%
+Coar.	48.6%	+IF	73.6%	+F	46.7%
+PS	0.4%	+PS	1.9%	+PS	0.6%

Table 2: RMSEs of syllable logF0 contour (**sp**), duration (*sd*), energy level (*se*), and pause duration (*pd*) for different size (number of syllable, syl #) of the adaptation data

syl #	**sp** ($\times 10^{-4}$(*Log-Hz*)^2)	*sd* (ms^2)	*se*(dB^2)	*pd*(ms^2)
10559	[0.80,10.0,4.76,1.92]	.321	0.33	55.7
9386	[0.80,10.0,4.64,1.87]	.331	0.34	56.0
7029	[0.81,10.0,4.62,1.92]	.335	0.35	57.9

Table 3: RMSE (*ms*) of reconstructed Pause Duration

	B0	B1	B2-1	B2-2	B2-3	B3	B4	avg
SMAP	0.9	10	6.5	27.3	5.6	101	147	40.9
ML	2.4	18.5	24.9	86.3	30.8	100	147	59.9

6. An Application to TTS

To examine the effectiveness of the proposed method, we use the prosody generated by the MAPLR NFs and the adaptive SR-HPM to synthesize a new voice by an HMM-based synthesizer [8]. The prosody generated by the MSE-estimated NFs and the ML-estimated SR-HPM [14] were taken as baselines for comparison. It is found from Table 4 that the RMSEs made by the proposed MAP-based approach were generally smaller than the ones by the conventional ML-based approach [14]. This result confirms the effectiveness of the proposed method objectively. The conventional system only performed slightly better than the proposed one for **sp** in the SR range 0.22~0.30 where most adaptation data lay in. This result is reasonable since the conventional method [14] could model the data laying in the seen SR range well, but might give poor estimation for data in unseen SRs.

To evaluate the proposed method subjectively, 15 subjects were involved in subjective tests to examine the naturalness of the synthesized speech. They were all graduate students of NCTU. 12 short paragraphs with length from 24 to 45 syllables were selected from the outside test data set. They were asked to give MOS and preference scores to the two utterances of each paragraph synthesized by the proposed system and the HTS system. As shown in Table 5, the MOS scores of the proposed method were better than the HTS method which simply realizes the SR effect by setting elastic factor ρ to the desired SR [4,8,14]. The SMAP method is slightly better for SR=0.22 which is in the observed SR range of target speaker's adaptation data, while it is much better for SR=0.3 which is in the unobserved SR range. These results confirmed the effectiveness of the proposed speaker adaptation method.

Table 4: RMSEs at different SRs on the test set. Note that numbers in brackets represent RMSE by the ML-estimated SR-HPM proposed in [14].

SR (*sec/syl*)	0.15~0.18	0.22~0.30	0.30~0.43
sp (log-Hz)	0.08 (0.17)	0.19 (0.18)	0.20 (0.36)
sd (*ms*)	49 (514)	68 (120)	96 (121)
pd (*ms*)	70 (78)	141 (148)	161 (223)

Table 5: The Results of MOS Test.

SR (*sec/syl*)		FAST(.15)	NORMAL (.22)	SLOW(.3)
MOS	SMAP	3.4	3.6	3.6
	ML	2.6	3.4	2.8
	HTS	3.2	3.4	3.0

7. Conclusions

An SMAP-based speaker adaptation method to adjust the existing SR-HPM of an SR-controlled Mandarin TTS system to new speaker's data for realizing a new voice of any given speaking rate has been discussed. The prior information provided by the reference SR-HPM is useful in the training of the adaptive SR-HPM to assist in solving the problems of data sparseness and model parameter extrapolation. Experimental results confirmed that the proposed method is promising.

8. Acknowledgements

This work was mainly supported by the MOST of Taiwan under Contract "MOST 103-2221-E-009-077-MY3" and partially under Contract "NSC-102-2221-E-305-005-MY3". The authors also would like to thank the ACLCLP for providing the Treebank Corpus.

9. References

[1] P. C. Wang, I. B. Liao, C. Y. Chiang, Y. R. Wang, S. H. Chen, "Speaker adaptation of speaking rate-dependent hierarchical prosodic model for Mandarin TTS," in *Proc. ISCSLP'14*, Singapore, Sept., 2014, pp.511-515.

[2] K.S.,C-H Lee, "A Structural Bayes Approach to Speaker Adaptation," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 3, MARCH 2001

[3] Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T., "Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR",in Proc. 3rd ESCA/COCOSDA Workshop Speech Synthesis, 273-276, Nov 1998.

[4] Hsieh, C. H., Wang, Y. R., Chiang, C. Y. and Chen, S. H., "A Speaking Rate-controlled Mandarin TTS System", in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 6900-6904, May 2013.

[5] Hsieh, C. H., Chiang, C. Y., Wang, Y. R., Yu, H. M. and Chen, S. H., "A New Approach of Speaking Rate Modeling for Mandarin Speech Prosody",in Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.(INTERSPEECH), 655-658, Sep 2012.

[6] Gauvain, J. L. and Lee C. H., "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. Speech and Audio Process.,2(2):291-298, Apr 1994.

[7] Leggetter,C. J. and Woodland, P. C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, 9:171-185, 1995.

[8] HTS Working Group, HTS-2.2 source code and demonstrations: http://hts.sp.nitech.ac.jp/?Download, accessed on Jul2014.

[9] Chen, S. H.and Wang,Y. R., "Vector Quantization of Pitch Information in Mandarin Speech", IEEE Trans. Commun., 38(9):1317-1320, Sep 1990.

[10] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., Classification and Regression Tree, Chapman and Hall/CRC, 1984.

[11] Chiang, C. Y., Chen, S. H., Yu, H. M. and Wang, Y. R., "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech", J. Acoust. Soc. Amer., 125(2):1164-1183, Feb 2009.

[12] Bishop, M., Pattern Recognition and Machine Learning, Springer Science, New York, 2009.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Belmont, CA, USA: Wadsworth, 1984.

[14] Sin-Horng Chen, Chiao-Hua Hsieh, Chen-Yu Chiang, Hsi-Chun Hsiao, Yih-Ru Wang, Yuan-Fu Liao, and Hsiu-Min Yu, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," in *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* , vol.22, no.7, pp.1158-1171, July 2014