Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

On Cross-Dialect and -Speaker Adaptation of Speaking Rate-Dependent Hierarchical Prosodic Model for a Hakka Text-to-Speech System

Chen-Yu Chiang [1], *Hsiu-Min Yu* [2], *Sin-Horng Chen* [3]

[1]Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan
[2]Language Center, Chung Hua University, Hsinchu, Taiwan
[3]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
cychiang@mail.ntpu.edu.tw, kuo@chu.edu.tw, schen@mail.nctu.edu.tw

Abstract

This paper presents an effective adaptation of an existing speaking rate-dependent hierarchical prosodic model (SR-HPM) for Mandarin to construct the SR-HPM for Hakka, another Chinese dialect. Based on the cross-dialectal linguistic similarities in terms of syntactic and prosodic structures, the adaptation is formulated as a maximum a posteriori estimation (MAP) problem with the existing Mandarin SR-HPM serving as an informative prior. In addition, benefiting from the well-trained Mandarin SR-HPM that models the effects of speaking rate (SR) on prosodic-acoustic features, the SR-HPM developed for Hakka could generate satisfactory prosody in various SRs. The performance of the approach proposed in this study was evaluated by an experiment of prosody generation for a SR-controlled Hakka text-to-speech system, in which the Hakka SR-HPM is trained by a Hakka corpus that is small in size and read in narrow SR. Results show that the generated Hakka prosody was judged to be quite natural by native Hakka speakers for SR varying from 3.3 syllables/sec to 6.7 syllables/sec.

Index Terms: prosody generation, dialects with limited size of corpus, Chinese, Mandarin, Hakka, text-to-speech systems

1. Introduction

Chinese, a term used as a language family and in its broad sense, is conventionally classified into seven dialect groups, namely, Guan, Wu, Yue, Min, Xiang, Hakka and Gan, ordered by decreasing number of dialect users within the group [1]. It is estimated that approximately 1.2 billion people (around 16% of the world's population) use some form of Chinese as their first language, and the numbers of native speakers in the Chinese dialects range from 31 million (for Gan) to 881 million (for Guan) [2]. With these large numbers of native speakers of each of the Chinese dialects respectively in mind, it is worthwhile to construct for each of the dialects a text-to-speech (TTS) system with an informative prosody modeling. It is even more tempting and practical to develop a cross-dialectal prosody modeling which can function as the prosody basis to the construction of TTS systems for all the Chinese dialects, given their shared linguistic characteristics in terms of, for example, the use of tonal features for lexical purposes, the uniform cross-dialectal syllabic structure [3], and a relatively high degree of similarity in the syntactic and discourse structures revealed in written or formal languages among Chinese dialects. Therefore, the purpose of this study is to present an effective adaptation of an existing Mandarin prosody modeling to help construct the prosodic model for a TTS system of Si-Xian Hakka, a sub-dialect Hakka used in Taiwan. To this end, the cross-dialect and -speaker prosody adaptation method is formulated based on a statistical Mandarin prosodic model, i.e. the speaking rate-dependent hierarchical prosodic model (SR-HPM), proposed in [4].

Fig. 1 shows an overview of the proposed approach/framework. The construction of the Mandarin SR-HPM and the adaptation for the Hakka SR-HPM are shown respectively in the upper and lower parts separated by the dashed line. Firstly, the training of the Mandarin SR-HPM starts with the construction of SR normalization functions (NFs) to obtain smooth normalization parameters for the following SR normalizations of the prosodic-acoustic features (PAFs). The use of the SR NFs is to compensate the effects of SR on PAFs. Then, the modified prosody labeling and modeling (PLM) algorithm [4] is conducted to simultaneously label a speech corpus with prosody tags of break types and prosodic states and to train the Mandarin SR-HPM. Here, the prosody tags are used to represent a four-layered prosodic structure, from bottom to top, including layers of syllable, prosodic word, prosodic phrase, and breath group/prosodic phrase, all defined in prosodic terms [4-6]. The main purpose of the modified PLM algorithm is to obtain prosodic labels of the speech corpus and to construct SR-HPM mainly by the observed PAFs with the help of linguistic features. Finally, a model refinement is made to increase the capability of the SR-HPM in prosody generation exclusively by using linguistic features.

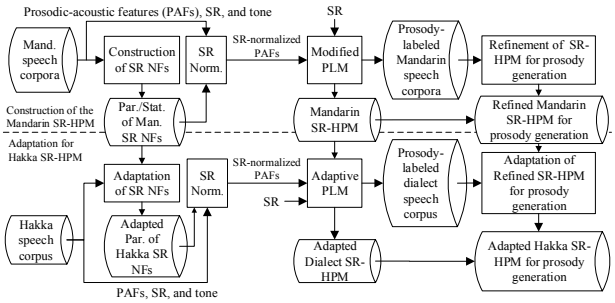

Fig. 1. Overview of the proposed framework of cross-dialect and –speaker SR-HPM adaptation.

The cross-dialect and -speaker adaptation for the Hakka SR-HPM operates in a similar flow like the training of the original Mandarin SR-HPM but in an adaptive fashion. It starts with the adaptation of the SR NFs. The prior probabilities in the adaptation processes are obtained by the statistics or parameters of Mandarin SR NFs. The parameters of the SR NFs for Mandarin provide a good reference for estimating SR NFs for Si-Xian Hakka since the SR coverage

10.21437/SpeechProsody.2016-161

of Mandarin corpora is larger than the SR coverage of Si-Xian Hakka corpus. After the adaptation of the parameters for SR NFs of Si-Xian Hakka, the adapted SR NFs are utilized to compensate the effects of SR on the PAFs of the dialect. Then, an adaptive PLM algorithm modified from the previous study [7] is formulated based on MAP estimation, and designed to simultaneously label prosody tags of the Si-Xian Hakka speech corpus and adapt model parameters of the SR-HPM for Si-Xian Hakka from the existing Mandarin SR-HPM. At the end, the adaptation of the refined SR-HPM is conducted to obtain the Si-Xian Hakka SR-HPM for prosody generation.

2. A General Prosody Generation Framework

Fig. 2 shows the general prosody generation framework which is originally designed for a SR-controlled Mandarin TTS [4]. In virtues of its flexibility and systematic module designs, this framework has already been applied for prosody generation of Taiwan Min dialect [8] and is applied for Si-Xian Hakka in this study. The SR-controlled prosody generation is powered by the SR-HPM and the SR NFs. In this study, the parameters of SR-HPM, i.e. $\{\boldsymbol{\lambda}_{\mathbf{B}}, \boldsymbol{\lambda}_{\mathbf{P}}, \boldsymbol{\lambda}_{\mathbf{PL}}, \boldsymbol{\lambda}_{\mathbf{YZ}}, \boldsymbol{\lambda}_{\mathbf{X}}\}$, and the SR NFs for a dialect are adapted from the ones for Mandarin.

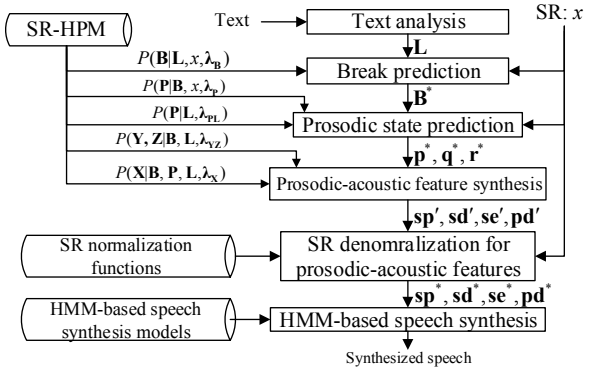

Fig. 2: Prosody generation powered by SR-HPM and SR NFs

First, the prosodic structure in terms of a break type sequence is generated by the SR-dependent break-syntax model, $P(\mathbf{B}|\mathbf{L}, \mathbf{x}, \boldsymbol{\lambda}_{\mathbf{B}})$, modeled by a decision tree, i.e.

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

where n stands for syllable index in an utterance; B_n and L_n represent respectively the break type for syllable juncture right after n-th syllable and contextual linguistic features for n-th syllable; x is the SR for the synthesis speech defined as the average number of syllables per second calculated with all pauses being excluded. It is noted that $P(\mathbf{B}|\mathbf{L}, \mathbf{x}, \boldsymbol{\lambda}_{\mathbf{B}})$ is implemented by a modified classification decision tree (DT) in which probability of each break type is a linear function of SR, x. Then, the global prosodic patterns in terms of 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, $P(\mathbf{P}|\mathbf{L}, \boldsymbol{\lambda}_{\mathbf{PL}}) = P(\mathbf{p}|\mathbf{L}, \boldsymbol{\lambda}_{\mathbf{pL}})P(\mathbf{q}|\mathbf{L}, \boldsymbol{\lambda}_{\mathbf{qL}})P(\mathbf{r}|\mathbf{L}, \boldsymbol{\lambda}_{\mathbf{rL}})$, given with the predicted prosodic structure (B_n^*), and the linguistic features ($\mathbf{L}$) and SR (x):

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

where $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{r}$ are prosodic state sequences for syllable logF0, syllable duration, and syllable energy level, respectively. The four SR-normalized PAFs 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, $P(\mathbf{Y},\mathbf{Z}|\mathbf{B}, \mathbf{L}, \boldsymbol{\lambda}_{\mathbf{YX}})$:

$$\begin{aligned} \mathbf{sp}'_n &= \boldsymbol{\beta}_{t_n} + \boldsymbol{\beta}_{p_n^*} + \boldsymbol{\beta}^{f}_{B^*_{n-1}, tp_{n-1}} + \boldsymbol{\beta}^{b}_{B^*_n, tp_n} + \boldsymbol{\mu}^{(sp)} \\ sd'_n &= \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n^*} + \mu^{(sd)} \\ se'_n &= \omega_{t_n} + \omega_{f_n} + \omega_{r_n^*} + \mu^{(se)} \\ pd'_n &= \kappa_{B^*_n, L_n}\, \theta_{B^*_n, L_n} \end{aligned} \tag{3}$$

where $\mathbf{sp}'_n$, sd'_n, se'_n, and pd'_n are respectively SR-normalized syllable logF0 contour, syllable duration, syllable energy level, and inter-syllable pause duration; $\boldsymbol{\beta}$'s, γ's, and ω'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 $tp_{n-1}=(t_{n-1}, t_n)$. The parameters κ's and Θ's are the parameters of the Gamma distribution for pause duration found from the leaf nodes of the DT of the syllable-juncture model given with the predicted break B_n^* and the linguistic features L_n. Last, to generate the PAFs affected by SR, the four PAFs are SR-denormalized by the inverse operation of the SR NFs:

$$\begin{aligned} \mathbf{sp}^*_n(i) &= \mathbf{sp}'_n(i) - \mu_g^{sp}(t_n, i) / \sigma_g^{sp}(t_n, i) \cdot \hat{\sigma}^{sp}(x, t_n, i) + \hat{\mu}^{sp}(x, t_n, i) \\ sd^*_n &= (sd'_n - \mu_g^{sd}) / \sigma_g^{sd} \cdot \hat{\sigma}^{sd}(x) + x \\ se^*_n &= se'_n \\ pd^*_n &= G^{-1}(G(pd'_n; \kappa_g^{pd}, \theta_g^{pd}); \hat{\kappa}^{pd}(x), \hat{\theta}^{pd}(x)) \end{aligned} \tag{4}$$

where $\mu_g^{sp}(t,i)$ and $\sigma_g^{sp}(t,i)$ are the global mean and standard deviation for the i-th dimension of tone t; $\hat{\mu}^{sp}(x,t,i)$ and $\hat{\sigma}^{sp}(x,t,i)$ are tone-dependent NFs for the i-th log-F0 component; $\{\mu_g^{sd}, \sigma_g^{sd}\}$ and $\{\kappa_g^{sd}, \Theta_g^{sd}\}$ are parameters representing the distributions of the SR-normalized syllable duration and pause duration, respectively; $\hat{\sigma}^{sd}(x)$ and $\{\hat{\kappa}^{pd}(x), \hat{\Theta}^{pd}(x)\}$ are the NFs for syllable duration standard deviation and pause duration. It is noted that $\hat{\mu}^{sp}(x,t,i)$ and $\hat{\sigma}^{sp}(x,t,i)$ are two 1st order polynomial functions of SR, x, while $\hat{\sigma}^{sd}(x)$ is a 2nd order polynomial function of SR, x. For modeling convenience, $\{\hat{\kappa}^{pd}(x), \hat{\Theta}^{pd}(x)\}$ are converted to the pause mean $\hat{\mu}^{pd}(x)$ and pause standard deviation $\hat{\sigma}^{pd}(x)$, which are modeled by two 2nd order polynomial functions of x.

3. Adaptation of Normalization Functions

3.1. Adaptation of Dialect-Independent SR NFs

Since the two types of NFs, i.e. $\hat{\sigma}^{sd}(x)$ and $\{\hat{\mu}^{pd}(x), \hat{\sigma}^{pd}(x)\}$ 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 mathematic formula for the adaptation of the syllable duration NF, $\hat{\sigma}^{sd}(x)$. Here, a variable $x(k)$ representing the average syllable duration of the k-th utterance is defined as an independent variable for the NF, $\hat{\sigma}^{sd}(\cdot)$. In the previous study [7], we found that the smoothed NF passing through the point of the average syllable standard deviation at the average SR of the target speaker corpus is preferable. Therefore, the parameters for $\hat{\sigma}^{sd}(x)$ are obtained by the following objective function of the MAPLR with a Lagrange multiplier λ:

$$\begin{aligned} a_0^*, a_1^*, a_2^* &= \arg\max_{a,b,c} \left[\ln P(a_0, a_1, a_2 | \boldsymbol{\sigma}^{sd}) + \lambda(\bar{\sigma} - a_0 - a_1\bar{x} - a_2\bar{x}^2) \right] \\ &\approx \arg\max_{a,b,c} \begin{bmatrix} \ln\left(P(\boldsymbol{\sigma}^{sd} | a_0, a_1, a_2)^{w(\mathbf{x})} P(a_0) P(a_1) P(a_2) \right) \\ + \lambda(\bar{\sigma} - a_0 - a_1\bar{x} - a_2\bar{x}^2) \end{bmatrix} \end{aligned} \tag{5}$$

where a's are parameters of the 2nd order polynomial; $P(\boldsymbol{\sigma}^{sd} | a_0, a_1, a_2)$ 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; $P(a_i)$ for i=0~2 is the prior probability of the i-th polynomial coefficient; $\bar{x}$ and $\bar{\sigma}$ are respectively the average SR and the average syllable-duration standard deviation of the target dialect corpus. The likelihood function is elaborated by

$$P(\boldsymbol{\sigma}^{sd} \mid a_0, a_1, a_2) = \prod_k N(\sigma^{sd}(k); \hat{\sigma}^{sd}(x(k)), v^{sd}) \quad (6)$$

where $\hat{\sigma}^{sd}(x(k))$ is the smooth NF modeled by a 2nd order polynomial:

$$\hat{\sigma}^{sd}(x) = a_0 + a_1 x + a_2 x^2 \quad (7)$$

; v^{sd} is the variance for $\sigma^{sd}(k)$; $w(\mathbf{x})$ is defined by

$$w(\mathbf{x}) = std(x(k)) / std(\hat{x}(k)) \quad (8)$$

where $std(x(k))$ and $std(\hat{x}(k))$ are the standard deviations of the observed utterance-wise SR of the target dialect speech corpus and the reference Mandarin speech corpus, respectively. The priors, $P(a_i)$ for $i = 0\sim2$, are all assumed to be Gaussian distributed. The means and variances for the priors are estimated by n-fold sets of the Mandarin speech corpora. Similar to the idea of the NF for the syllable duration standard deviation, $\hat{\sigma}^{sd}(x)$, the parameters of the NFs for pause-duration mean, $\hat{\mu}^{pd}(x)$, and standard deviation, $\hat{\sigma}^{pd}(x)$ are also estimated by the MAPLR method, given with the observed utterance-wise pause duration means/standard deviations, and the associated priors estimated from the reference Mandarin speech corpora.

3.2. Adaptation of Dialect-Dependent SR NFs

Since the tones of a Chinese dialect are similar to the ones of Mandarin Chinese, we can still apply the logF0 contour NFs defined in Eq. (4) for a dialect. Note that the NFs, $\hat{\mu}^{sp}(x,t,i)$ and $\hat{\sigma}^{sp}(x,t,i)$ for $t \in$ {the set of dialectal tones} and $i = 0\sim3$, are both linear functions of the same forms. We can still apply the MAPLR method as illustrated in Section 3.1 to estimate their parameters. For simplicity, we only illustrate the estimation of parameters for $\hat{\mu}^{sp}(x,t,i)$. Again, a general objective function of the MAPLR is expressed by

$$\{a_0(t,i), a_1(t,i)\}^* = \arg\max_{a_0(t,i), a_1(t,i)} \left\{ \sum_k \ln \left[P\left(\mu^{sp}(k,t,i) \mid a_0(t,i), a_1(t,i)\right)^{w(\mathbf{x})} P\left(a_0(t,i)\right) P\left(a_1(t,i)\right) \right] + \lambda(\bar{\mu}^{sp}(t,i) - a_0(t,i) - a_1(t,i)\bar{x}) \right\} \quad (9)$$

where $P(\mu^{sp}(k,t,i)|a_0(t,i),a_1(t,i))$ is the likelihood function; $\mu^{sp}(k,t,i)\sim N(a_0(t,i)+a_1(t,i)x, v(t,i))$ is the average value of the i-th component of dialectal tone t for utterance k; $a_0(t,i)+a_1(t.i)x$ is the smooth logF0 mean for i-th dimension of dialectal tone t; $v(t,i)$ is the associated variance; $P(a_j(t,i))$ for j={0,1} are prior probabilities and $a_j(t,i)\sim N(\mu_{a_j(t,i)}, v_{a_j(t,i)})$ for j={0,1}; $\mu_{a_j(t,i)}$ and $v_{a_j(t,i)}$ are the prior mean and variance for the parameter $a_j(t,i)$; $\bar{\mu}^{sp}(t,i)$ is the average value of the whole dialect corpus for the i-th component of dialect tone t.

However, we do not have priors for the parameters of $\hat{\mu}^{sp}(x,t,i)$ and $\hat{\sigma}^{sp}(x,t,i)$ due to the fact that we lack of large dialect speech corpora of various SRs to obtain these informative priors. Fortunately, the characteristics of Chinese dialectal tones may be similar to the ones of Mandarin. Therefore, the priors of dialectal tones can be synthesized by the priors of Mandarin tones, i.e.

$$\begin{aligned} \mu_{a_j(t,i)} &= \sum_{u=1}^{5} S(t,u)\hat{\mu}_{a_j(u,i)} \\ v_{a_j(u,i)} &= \sum_{u=1}^{5}\left[S(t,u)\left(\hat{v}_{a_j(u,i)} + (\hat{\mu}_{a_j(u,i)})^2\right)\right] - (\mu_{a_j(t,i)})^2 \end{aligned} \quad (10)$$

where $S(t,u)$ is a similarity measure between Chinese dialect tone t and Mandarin tone u; $\hat{\mu}_{a_j(u,i)}$ and $\hat{v}_{a_j(u,i)}$ are the mean and variance of the prior for coefficient $a_j(u,i)$ of Mandarin tone u. The similarity measure $S(t,u)$ is defined by

$$S(t,u) = AL(t,u) / \sum_{h=1}^{5} AL(t,h) \quad (11)$$

where $AL(t,u)$ is defined as the average likelihood of the dialect tone t w.r.t. Mandarin tone u:

$$AL(t,u)=\exp\left(\sum_{k,n} \ln N(\mathbf{sp}''_{n,k}; \boldsymbol{\beta}_u, \boldsymbol{\sigma}_u^{sp})\delta(t_{n,k}=t) / \sum_{k,n}\delta(t_{n,k}=t)\right) \quad (12)$$

where $\mathbf{sp}''_{n,k}$ represents logF0 contour obtained after frame-based Gaussian normalization to the standard deviation of logF0 of Mandarin speech; $\boldsymbol{\beta}_u$ and $\boldsymbol{\sigma}_u^{sp}$ are mean pattern and standard deviation of Mandarin tone u.

4. Adaptation of the SR-HPM for Dialect

As shown in Fig. 1, after the adaptation of the SR NFs for a dialect, two main steps are executed to obtain the SR-HPM for dialect prosody generation: 1) the adaptive PLM algorithm and 2) the adaptation of the refined SR-HPM for prosody generation. Both steps are formulated based on MAP estimations. The main purpose of the adaptive PLM algorithm is to simultaneously obtain optimal prosody labels of a dialect speech corpus and adapt the SR-HPM for a dialect from the one for Mandarin mainly by the observed PAFs with the auxiliary of linguistic features. The adaptation of the refined SR-HPM for prosody generation is to adapt the enhanced sub-models for prosody generation for a dialect, i.e. $\hat{\boldsymbol{\lambda}}_{\mathbf{B}}$ and $\{\boldsymbol{\lambda}_{u\mathbf{L}}\}|u = \mathbf{p}, \mathbf{q}, \mathbf{r}$, from the ones for Mandarin.

Since Chinese dialects share similar linguistic characteristics with Mandarin, most model parameters of the Mandarin SR-HPM, except tone-related and base-syllable type-related parameters, can directly serve as the priors for the Hakka SR-HPM. The tone-related APs can be linear combinations of the Mandarin's APs with coefficients set to be the tone similarity measure defined in Eq. (11). For the dialect APs of base-syllable and final types, we cluster the associated Mandarin APs by a DT into several clusters, according to a question set made of phonetic features. Therefore, a proper prior for an AP of dialect's base-syllable/final type can be selected from a leaf node of the DT by traversing the branches according to the phonetic features of dialect's syllable/final.

4.1. The Adaptive PLM algorithm

The adaptive PLM algorithm is formulated based on the MAP estimation with the Mandarin SR-HPM serving as an informative prior. It is designed to simultaneously estimate the subset of the model parameters of the dialect SR-HPM, $\boldsymbol{\lambda}^* = \{\boldsymbol{\lambda}_{\mathbf{B}}, \boldsymbol{\lambda}_{\mathbf{P}}, \boldsymbol{\lambda}_{\mathbf{YZ}}, \boldsymbol{\lambda}_{\mathbf{X}}\}^*$, and label the prosody tags of target dialect, $\mathbf{T}^*$, given with PAFs, $\mathbf{A}$, linguistic features, $\mathbf{L}$, and SR, $\mathbf{x}$:

$$\boldsymbol{\lambda}^*, \mathbf{T}^* = \arg\max_{\boldsymbol{\lambda}, \mathbf{T}} P(\boldsymbol{\lambda}, \mathbf{T} | \mathbf{A}, \mathbf{L}, \mathbf{x}) = \arg\max_{\boldsymbol{\lambda}, \mathbf{T}} P(\mathbf{T}, \mathbf{A} | \mathbf{L}, \mathbf{x}, \boldsymbol{\lambda}) P(\boldsymbol{\lambda}) \quad (13)$$

where $P(\mathbf{T}, \mathbf{A}|\mathbf{L}, \mathbf{x}, \boldsymbol{\lambda})$ is the original SR-HPM in [4]; $P(\boldsymbol{\lambda}) = P(\boldsymbol{\lambda}_{\mathbf{B}})P(\boldsymbol{\lambda}_{\mathbf{P}})P(\boldsymbol{\lambda}_{\mathbf{YZ}})P(\boldsymbol{\lambda}_{\mathbf{X}})$ is the prior probability for the dialectal SR-HPM parameters. The adaptive PLM algorithm is specially designed for the 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 follows:

Step 1: Set all the parameters of SR-HPM as their prior means.
Step 2: Find the initial break type sequence by

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

Step 3: Obtain the optimal prosodic state sequences by

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} P(\mathbf{X}|\mathbf{B}^*,\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\mathbf{P}|\mathbf{B}^*,\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}}) \quad (15)$$

Step 4: Adapt the sets of $\boldsymbol{\lambda}_{\mathbf{X}}$, $\boldsymbol{\lambda}_{\mathbf{YZ}}$, $\boldsymbol{\lambda}_{\mathbf{B}}$, and $\boldsymbol{\lambda}_{\mathbf{P}}$ by the following MAP estimations:

$$\begin{aligned}
\boldsymbol{\lambda}_{\mathbf{X}}{}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{X}}} P(\mathbf{X}|\mathbf{B}^*,\mathbf{P}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\boldsymbol{\lambda}_{\mathbf{X}}) \\
\boldsymbol{\lambda}_{\mathbf{YZ}}{}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{YZ}}} P(\mathbf{Y},\mathbf{Z}|\mathbf{B}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{YZ}})P(\boldsymbol{\lambda}_{\mathbf{YZ}}) \\
\boldsymbol{\lambda}_{\mathbf{B}}{}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{B}}} P(\mathbf{B}^*|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}})P(\boldsymbol{\lambda}_{\mathbf{B}}) \\
\boldsymbol{\lambda}_{\mathbf{P}}{}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{P}}} P(\mathbf{P}|\mathbf{B}^*,\boldsymbol{\lambda}_{\mathbf{P}})P(\boldsymbol{\lambda}_{\mathbf{P}})
\end{aligned} \quad (16)$$

Step 5: Find the optimal break sequence by the Viterbi search:

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} \begin{bmatrix} P(\mathbf{X}|\mathbf{B},\mathbf{P}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}}{}^*)P(\mathbf{Y},\mathbf{Z}|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{YZ}}{}^*) \\ P(\mathbf{P}^*|\mathbf{B},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}}{}^*)P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}}{}^*) \end{bmatrix} \quad (17)$$

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. The Adaptation of the Refined SR-HPM

After conducting the adaptive PLM algorithm, we can obtain the optimal break types, $\mathbf{B}^*$, and the optimal prosodic-state tags, $\mathbf{P}^*=\{\mathbf{p}^*, \mathbf{q}^*, \mathbf{r}^*\}$. Then, given with $\mathbf{B}^*$, the associated linguistic features, $\mathbf{L}$, and the refined SR-dependent break-syntax model for Mandarin, $\hat{\boldsymbol{\lambda}}_{\mathbf{B}}$, the refined SR-dependent break-syntax model for a dialect, $\boldsymbol{\lambda}_{\mathbf{B}}^*$, can be adapted by

$$\boldsymbol{\lambda}_{\mathbf{B}}{}^* = \arg\max_{\hat{\boldsymbol{\lambda}}_{\mathbf{B}}} P(\mathbf{B}^*|\mathbf{L},\mathbf{x},\hat{\boldsymbol{\lambda}}_{\mathbf{B}})P(\hat{\boldsymbol{\lambda}}_{\mathbf{B}}) \quad (18)$$

The three prosody-syntax models, $\{\boldsymbol{\lambda}_{u\mathbf{L}}\}|u = \mathbf{p},\mathbf{q},\mathbf{r}$, can also be adapted in the same fashion:

$$\boldsymbol{\lambda}_{u\mathbf{L}}{}^* = \arg\max_{\boldsymbol{\lambda}_{u\mathbf{L}}} P(u^*|\mathbf{L},\boldsymbol{\lambda}_{u\mathbf{L}})P(\boldsymbol{\lambda}_{u\mathbf{L}}) \quad \text{for } u = \mathbf{p},\mathbf{q},\mathbf{r} \quad (19)$$

5. Prosody Generation Experiment

The adaptation experiment is conducted on a male-speaker Si-Xian Hakka database [9] with 159 paragraphic utterances (15,009 syllables) and 39 paragraphic utterances (3,711 Syllables) for adaptation and testing, respectively. The Mandarin SR-HPM is trained by a female Mandarin speech database used in the previous studies [4]. The database contains four parallel speech corpora of slow, medium, normal and fast SRs. There are in total 1,478 paragraphic utterances with 183,795 syllables. The SR of the Mandarin database covers a wide range of 3.4-6.8 syl/sec while the SR of the Hakka corpus only covers a smaller range of 3.8-5.1 syl/sec. Table 1 shows the root mean squared errors (RMSEs) of the PAFs for the test set w.r.t. different adaptation data sizes. It is noted that all the PAFs were generated given with the target SRs of the testing utterances. The PAFs of the testing utterances were taken as ground truth. It is found that the RMSEs of most PAFs gradually decreased as the size of training/adaptation data increased for both the proposed MAP-estimated SR-HPM and the baseline maximum likelihood (ML)-estimated SR-HPM [4]. The RMSEs by the proposed MAP method were generally smaller than the ones by the conventional ML method. We may conclude that the results shown in Table 1 partially confirm the effectiveness of the proposed approaches of NF adaptation and adaptive SR-HPM for the dialect in this resource-limited condition.

Last, Mean Opinion Score (MOS) and preference tests were conducted to compare the quality of the paragraphic utterances produced by the HMM-based synthesizer that is controlled by the prosody generated by the proposed MAP-estimated SR-HPM with that of utterances controlled by the conventional ML-estimated SR-HPM [4]. The HMM-based speech synthesis [10,11] was constructed by the HTS-2.2 toolkit with all the utterances in the adaptation set (15,009 syllables). Sub-syllable units of initial and final were taken as basic HMM synthesis unit [4]. The ML-estimated and MAP-estimated SR-HPMs were trained or adapted by the training/adaptation data with 15,009 syllables. Three native Hakka speakers and one much experienced Hakka subject were recruited to participate in the subjective tests and each participant was asked to listen to four synthesized short Si-Xian Hakka paragraphs for each of the eight SRs and for each of the two TTS systems. Totally, this subjective listening assessment experiment comprised 256 trials (4 paragraphs x 8 SRs x 2 TTS systems x 4 participants). Before listening to the synthesized utterances, subjects were asked to listen to the authentic utterances in the test speech corpus corresponding to the synthesized speeches as a reference for their assessment opinion. Table 2 shows the results of the subjective tests for eight SRs. Table 2 indicates that both the MOSs and the preferences by the proposed MAP-estimated SR-HPM are generally higher than the ones by the baseline ML-estimated SR-HPM in a wide range of SR except for the cases in 4.35 syl/sec. This exception may be understandable because the average SR of the adaptation speech corpus is around 4.35 syl/sec, resulting in a good ML-estimated SR-HPM for Hakka. Notwithstanding of this exception, the proposed method still could generate more natural speech prosody than the baseline ML method [4] in the extrapolated (unseen) SR ranges of 6.70-5.26 syl/sec and 4.00-3.33 syl/sec. These results partially confirm the effectiveness of the proposed method.

Table 1. *RMSEs of logF0 contour (sp), syllable duration (sd), syllable energy level (se), and pause duration (pd) w.r.t. different size (number of syllable, syl#) of the adaptation data.*

	The proposed method (MAP)				The baseline (ML)			
syl#	*sp*(logHz)	*sd*(ms)	*se*(dB)	*pd*(ms)	*sp*	*sd*	*se*	*pd*
2,266	.189	70.7	5.57	101	.201	77.6	5.65	91
6,019	.183	69.2	5.49	101	.190	74.6	5.20	141
8,720	.182	69.4	5.49	95	.184	79.6	5.28	107
12,127	.181	69.7	5.49	93	.183	76.4	5.28	128
15,009	.179	69.8	5.48	98	.183	71.2	5.52	120

Table 2: *The results of the MOS and the preference tests.*

SR(syl/sec)=1/*x*		6.70	5.88	5.26	4.76	4.35	4.00	3.57	3.33
MOS	MAP	2.00	2.44	2.94	3.19	2.94	3.13	3.25	2.75
	ML	1.69	1.88	2.19	2.94	3.06	3.00	2.06	1.69
Prefer. (%)	MAP	62.5	75.0	81.3	50.0	18.8	37.5	93.8	93.8
	ML	18.8	12.5	6.3	18.8	37.5	31.3	6.3	6.3
	equal	18.8	12.5	12.5	31.3	43.8	31.3	0.0	0.0

6. Conclusions

This paper has proposed an MAP-based cross-dialect and -speaker adaptation approach to constructing the Si-Xian Hakka SR-HPM for prosody generation in a SR-controlled Hakka TTS. The adapted Hakka SR-HPM can generate quite natural prosody for a wide range of SR. Up to now, the SR-HPM framework has been successfully applied to Chinese dialects of Mandarin (Guan), Taiwanese (Min), and Si-Xian Hakka. We believe the proposed SR-HPM framework could be further applied to the other Chinese dialects in the future.

7. Acknowledgements

This work was mainly supported by the MOST of Taiwan under Contract No. NSC-102-2221-E-305-005-MY3 and partially under Contract MOST-103-2221-E-009-077-MY3.

8. References

[1] Yuan, Jiahua, et al. *Hanyu fangyan gaiyao* 汉语方言概要. 2nd ed. Beijing: Yuwen Chubanshe, 2001. (in Chinese)

[2] https://en.wikipedia.org/wiki/Chinese_language

[3] Duanmu, San (1990) A Formal Study of Syllable, Tone, Stress and Domain in Chinese Languages. Doctoral dissertation. Cambridge, MA: MIT.

[4] S.-H. Chen, C.-H. Hsieh, C.-Y. Chiang, H.-C. Hsiao, Y.-R. Wang, Y.-F. Liao, and H.-M. Yu, "Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 22, Issue 7, 1158 - 1171, 2014.

[5] 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.

[6] C.-Y. Tseng, S.-H. Pin, Y.-L. Lee, H.-M. Wang, and Y.-C. Chen, "Fluent speech prosody: framework and modeling," *Speech Communication*, Vol. 46, Issues 3-4, Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation, 284-309

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

[8] C.-Y. Chiang, "A Study on Adaptation of Speaking Rate-Dependent Hierarchical Prosodic Model for Chinese Dialect TTS," accepted by *O-COCOSDA 2015*.

[9] Y.-L. Tsai, H.-M. Yu, Y.-R. Wang, C.-Y. Chiang, L.-S. and Lo, S.-H. Chen, "An HMM-based Hakka Text-to-Speech System," in *Proc. O-COCOSDA 2010*, Nepal, Oct. 2010.

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

[11] The HTS working group, HTS-2.2 source code and demonstrations, available: http://hts.sp.nitech.ac.jp/?Download