SPEAKER ADAPTATION FOR SPEAKING RATE-CONTROLLED MANDARIN TTS

I-Bin Liao[1, 2], Advisor:*Sin-Horng Chen*[2]

[1]TL of Chunghwa Telecom Co., Ltd
[2]ECE Department of National Chiao Tung University, Hsinchu , Taiwan
snet@cht.com.tw, schen@mail.nctu.edu.tw

Abstract

In this research, a structural maximum a posteriori (SMAP) speaker adaptation method to adpat Speaking Rate dependent Hierarchical Prosodic Model (SR-HPM) for generating a Personalized SR-TTS 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. Spectrum model is bulit via Speaker Adaptation Training which incorporates an SMAP criterion which uses tree structures of the distributions to effectively cope with the control of the hyperparameters. Combining context-dependent linguistic features and prosody-dependent features generated by adaptive SR-HPM for personalized-spectrum modeling. 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-Controlled Mandarin TTS system (SR-MTTS) using a large speech corpus containing utterances of various SRs (0.15~0.3 sec/syl) of a female speaker. An SR-HPM was trained and used in the TTS system to generate prosodic-acoustic features for any given SR. In order to construct a personalized TTS system , two issues are 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.

Since the speech of a speaker can be generally characterized by its spectrum and prosody. Many speaker adaptation techniques have been proposed in the past. Among them, maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) are two popular approaches for spectral model adaptation. Actually, there are very few literatures relating to speaker prosody adaptation. This is owing to the scarcity of sophisticated prosodic models used in TTS. Another approach, MAP-based prosody adaptation method, was reported to adapt the condition random field (CRF)-based break prediction model of a source speaker to one for a target speaker. Its prosody adaptation is only fair because of using simple human-labeled break tags to adjust the decision tree, which generates log-F0 and state duration, without the help of any prosodic model. In this research, a structural MAP (SMAP) adaptation method is proposed to tackle these two issues. It organizes the model parameters of the SR-HPM into hierarchical structures so as to effectively perform an MAP-based speaker adaptation given with a new speaker's dataset with utterances covering a narrow SR range.

2. Overview of Personalized TTS system

Fig.1 displays a block diagram of Personalized TTS system. Using an input text and a given speaking rate x, the system first predicts the break sequence $\mathbf{B}^*$ by using the break-syntax sub-model $P(\mathbf{B}|\mathbf{L},x,\boldsymbol{\lambda}_\mathbf{B})$ of the adaptive SR-HPM. It then predicts the three prosodic state sequences ($\mathbf{p}^*,\mathbf{q}^*,\mathbf{r}^*$) by using the prosodic state sub-model $P(\mathbf{P}|\mathbf{B},x,\boldsymbol{\lambda}_\mathbf{P})$ and the prosodic state-syntax sub-model $P(\mathbf{P}|\mathbf{L},\boldsymbol{\lambda}_\mathbf{PL})$ of the adaptive SR-HPM. Then, it uses $\mathbf{B}^*$, ($\mathbf{p}^*,\mathbf{q}^*,\mathbf{r}^*$), x, and linguistic features L to generate the four prosodic-acoustic feature sequences $\{\mathbf{sp}^*,\mathbf{sd}^*,\mathbf{se}^*,\mathbf{pd}^*\}$ by using the syllable prosodic-acoustic sub-model $P(\mathbf{X}'|\mathbf{B},\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_\mathbf{X})$, the syllable juncture prosodic-acoustic sub-model $P(\mathbf{Y}',\mathbf{Z}'|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_\mathbf{YZ})$ and denormalization functions (inverse NFs) of the adaptive SR-HPM. Lastly, it produces the synthetic speech using the spectral features, generated by the constrained structural maximum a posteriori linear regression algorithm (CSMAPLR), and the prosodic-acoustic features $\mathbf{A}=\{\mathbf{sp}^*,\mathbf{sd}^*,\mathbf{se}^*,\mathbf{pd}^*\}$.

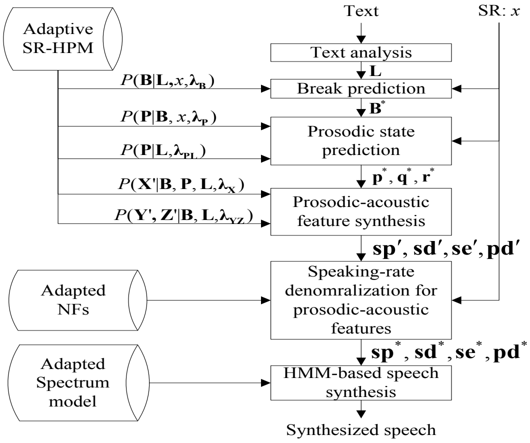
Adaptive SR-HPM
Text
SR: x
Text analysis
$\mathbf{L}$
$P(\mathbf{B}|\mathbf{L},x,\boldsymbol{\lambda}_\mathbf{B})$
Break prediction
$\mathbf{B}^*$
$P(\mathbf{P}|\mathbf{B},x,\boldsymbol{\lambda}_\mathbf{P})$
$P(\mathbf{P}|\mathbf{L},\boldsymbol{\lambda}_\mathbf{PL})$
Prosodic state prediction
$\mathbf{p}^*,\mathbf{q}^*,\mathbf{r}^*$
$P(\mathbf{X}'|\mathbf{B},\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_\mathbf{X})$
$P(\mathbf{Y}',\mathbf{Z}'|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_\mathbf{YZ})$
Prosodic-acoustic feature synthesis
$\mathbf{sp}',\mathbf{sd}',\mathbf{se}',\mathbf{pd}'$
Adapted NFs
Speaking-rate denomralization for prosodic-acoustic features
$\mathbf{sp}^*,\mathbf{sd}^*,\mathbf{se}^*,\mathbf{pd}^*$
Adapted Spectrum model
HMM-based speech synthesis
Synthesized speech

Fig. 1. A block diagram of Personalized TTS system.

The speaker adaptation of SR-HPM starts with the adaptation of NFs by the MAP linear regression (MAPLR) method using the NFs of the reference speaker (REF NFs), and the prosodic-acoustic features **A** and speaking rates **x** of the target speaker's speech corpus. **A** is then normalized using the adapted NFs. Then, the adaptation of SR-HPM is performed by the Adaptive Prosody Labeling and Modeling (A-PLM) algorithm using the normalized features $\mathbf{A}'=\{\mathbf{sp}',\mathbf{sd}',\mathbf{se}',\mathbf{pd}'\}$ and the REF SR-HPM. The structure MAP (SMAP) method is employed in the A-PLM algorithm to adapt parameters of sub-models.

3. 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, $\mathbf{x}$:

$$\begin{aligned}\boldsymbol{\lambda}^*,\mathbf{T}^* &= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\boldsymbol{\lambda}\,|\,\mathbf{T},\mathbf{A}',\mathbf{L},\mathbf{x}) = \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\boldsymbol{\lambda},\mathbf{T},\mathbf{A}',\mathbf{L},\mathbf{x}) \\ &= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\mathbf{T},\mathbf{A}'\,|\,\mathbf{L},\mathbf{x},\boldsymbol{\lambda})P(\mathbf{L},\mathbf{x}\,|\,\boldsymbol{\lambda})P(\boldsymbol{\lambda}) \\ &= \arg\max_{\boldsymbol{\lambda},\mathbf{T}} P(\mathbf{T},\mathbf{A}'\,|\,\mathbf{L},\mathbf{x},\boldsymbol{\lambda})P(\boldsymbol{\lambda})\end{aligned} \tag{1}$$

where $P(\mathbf{T},\mathbf{A}'|\mathbf{L},\mathbf{x},\boldsymbol{\lambda})$ is likelihood function and $P(\boldsymbol{\lambda})$ is the prior probability for the SR-HPM parameters.

3.1. The Adaptive PLM algorithm

The A-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

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} P(\mathbf{Y}',\mathbf{Z}'|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y},\mathbf{Z}})P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}}) \tag{2}$$

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

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} P(\mathbf{X}'|\mathbf{B}^*,\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\mathbf{P}|\mathbf{B}^*,\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}}) \tag{3}$$

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:

$$\begin{aligned}\boldsymbol{\lambda}_{\mathbf{X}}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{X}}} P(\mathbf{X}'|\mathbf{B}^*,\mathbf{P}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\boldsymbol{\lambda}_{\mathbf{X}}) \\ \boldsymbol{\lambda}_{\mathbf{Y,Z}}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{Y,Z}}} P(\mathbf{Y}',\mathbf{Z}'|\mathbf{B}^*,\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y,Z}})P(\boldsymbol{\lambda}_{\mathbf{Y,Z}}) \\ \boldsymbol{\lambda}_{\mathbf{B}}^* &= \arg\max_{\boldsymbol{\lambda}_{\mathbf{B}}} P(\mathbf{B}^*|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}})P(\boldsymbol{\lambda}_{\mathbf{B}}) \\ \boldsymbol{\lambda}_{\mathbf{P}}^* &= \arg\max_{\boldsymbol{\lambda}_{P}} P(\mathbf{P}|\mathbf{B}^*,\boldsymbol{\lambda}_{\mathbf{P}})P(\boldsymbol{\lambda}_{\mathbf{P}})\end{aligned} \tag{4}$$

$\boldsymbol{\lambda}_{\mathbf{X}} = \boldsymbol{\lambda}_{\mathbf{X}}^*$, $\boldsymbol{\lambda}_{\mathbf{YZ}} = \boldsymbol{\lambda}_{\mathbf{YZ}}^*$, $\boldsymbol{\lambda}_{\mathbf{B}} = \boldsymbol{\lambda}_{\mathbf{B}}^*$, and $\boldsymbol{\lambda}_{\mathbf{P}} = \boldsymbol{\lambda}_{\mathbf{P}}^*$

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

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} \begin{bmatrix} P(\mathbf{X}'|\mathbf{B},\mathbf{P},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{X}})P(\mathbf{Y}',\mathbf{Z}'|\mathbf{B},\mathbf{L},\boldsymbol{\lambda}_{\mathbf{Y,Z}}) \\ P(\mathbf{P}|\mathbf{B},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{P}})P(\mathbf{B}|\mathbf{L},\mathbf{x},\boldsymbol{\lambda}_{\mathbf{B}}) \end{bmatrix} \tag{5}$$

and update break type tags by $\mathbf{B} = \mathbf{B}^*$

Step 6: If convergent, then go to **Step 7**; or go to **Step 3**.

Step 7: Adapt the prosodic state-syntax sub-model $\boldsymbol{\lambda}_{\mathbf{PL}}$ by

$$\boldsymbol{\lambda}_{\mathbf{PL}}^* = \arg\max_{\boldsymbol{\lambda}_{\mathbf{PL}}} P(\mathbf{P}|\mathbf{L},\boldsymbol{\lambda}_{\mathbf{PL}})P(\boldsymbol{\lambda}_{\mathbf{PL}})$$

4. Experimental Results

Effectiveness of the proposed method was examined by simulations on the speech corpus of a female target speaker. The adaptation data contained 37 paragraphic utterances with 4,268 syllables uttered in a range of slow SR, i.e. 0.26~0.3 sec/syl. The testing corpus contained 12 utterances with 1,461 syllables. The average length of these utterances was 117 syllables.

4.1. Prosodic/Spectral Feature Prediction

Table I shows the root mean squared errors (RMSEs) of the PAFs w.r.t. different adaptation data sizes. It is found that the RMSEs for the four types of PAFs are approximately the same for the proposed SMAP method, while they decrease gradually as the adaptation data size increases for the ML method. As seen from Table II, PD labels is more efficiently for spectral feature generation than CD linguistic features. MCDs of 6.137, 5.905, 5.910, 5.766, 5.734 and 5.477 are achieved by the CSMAPLR method using PD features, while they are 6.336, 5.909, 5.997, 6.051, 5.972 and 5.822 for the case of using CD features.

TABLE I: PROSODY PREDICTION ERRORS (RMSES) OF FOUR PROSODIC-ACOUSTIC FEATURES FOR THE ADAPTIVE SR-HPM TRAINED BY USING ADAPTATION DATA OF 5 DIFFERENT SIZES. THE TEST DATA SIZE IS 1,461 SYLLABLES.

Spk.	Female (test data syl# 1,461)							
	The proposed SMAP method				ML-based HTS method			
syl#	**sp** (logHz)	*sd* (ms)	*se* (dB)	*pd* (ms)	**sp** (logHz)	*sd* (ms)	*se* (dB)	*pd* (ms)
1,417	.173	8.6	3.23	11	.229	11.4	4.23	34
2,000	.171	8.6	3.25	10	.216	10.1	4.36	18
2,661	.172	8.5	3.23	11	.224	9.8	4.27	16
3,348	.171	8.4	3.22	11	.206	9.3	3.38	11
4,268	.170	8.4	3.24	10	.213	8.6	3.3	11

TABLE II : THE MEL-CEPSTRAL DISTORTION (MCD) OF SPECTRAL FEATURE FOR DIFFERENT CONTEXT LABELS ; (UNIT: DB)

syl#	Context-Dependent linguistic features	Prosody- Dependent features
1,417	6.336	6.137
2,661	5.997	5.910
4,268	5.822	5.477

4.2. Subjective Test

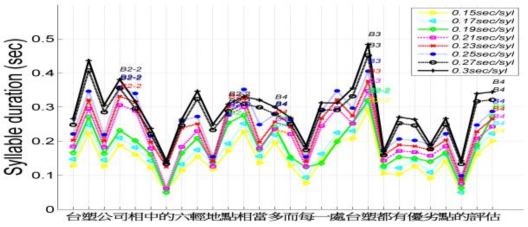

Fig. 2. An example of break type predictions and their pause durations for 8 SRs by the Personalized TTS.

Fig. 2. shows the syllable duration estimation. As shown in the figure, syllable durations are larger before short- and long-pause breaks(B2-2/B3/B4) to demonstrate the pre-boundary lengthening effect. These results match with the prior knowledge about the relationship between syllable juncture break pause and speaking rate.

TABLE III: EXPERIMENTAL RESULTS OF PREFERENCE TEST

SR (*sec/syl*)	0.15	0.23	0.3
Prefer SMAP (%)	70.5	77.0	92.7
Prefer ML (%)	10.8	18.8	7.3
Equal (%)	18.8	4.2	0.0

As shown in Table III , the average preference score (95% confidence intervals) of the SMAP method is 80.06 (78.6~81.5), while that of the ML method is 12.3 (11.55~13.04).

5. Conclusion

The proposed adaptation method simultaneously labels prosody tags on all utterances of the adaptation corpus and adapts model parameters of the SR-HPM for a target speaker with the help of the existing Mandarin SR-HPM. No human prosody labeling is needed. It requires no parallel speech corpora for speaker prosody adaptation and generate an adaptive SR-HPM to cover the whole SR range by using adaptation data that have a narrow SR coverage.