Interspeech 2018
2-6 September 2018, Hyderabad, India

An Exploration of Local Speaking Rate Variations in Mandarin Read Speech

Guan-Tin Liou[1], Chen-Yu Chiang[2], Yih-Ru Wang[1] and Sin-Horng Chen[1]

[1]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
[2]Dept. of Communication Engineering, National Taipei University, New Taipei City, Taiwan
tn00320663@gmail.com, cychiang@mail.ntpu.edu.tw, {yrwang, schen}@mail.nctu.edu.tw

Abstract

This paper explores speaking rate variation in Mandarin read speech. In contrast to assuming that each utterance is generated in a constant or global speaking rate, this study seeks to estimate local speaking rate for each prosodic unit in an utterance. The exploration is based on the existing speaking rate-dependent hierarchical prosodic model (SR-HPM). The main idea is to first use the SR-HPM to explore the prosodic structures of utterances and extract the prosodic units. Then, local speaking rate is estimated for each prosodic unit (prosodic phrase in this study). Some major influence factors including tone, base syllable type, prosodic structure, and speaking rate of the higher prosodic units (utterance and BG/PG) are compensated in the local SR estimation. A syntactic-local SR model is constructed and use in the prosody generation of Mandarin TTS. Experimental results on a large read speech corpus generated by a professional female announcer showed that the generated prosody with local speaking rate variations is proved to be more vivid than the one with a constant speaking rate.

Index Terms: speaking rate, SR-HPM, speech rate, articulation rate, prosody, text-to-speech, Mandarin

1. Introduction

Modeling speaking rate (SR) is useful for many speech applications, such as automatic speech recognition (ASR), emotion recognition, and text-to-speech system (TTS). For TTS, generating speech in a user-defined or controllable SR makes the synthesized speech more vivid and suitable for various applications, e.g., fast speech for visually-impaired people and slow speech for language learners. Many SR-modeling methods for TTS were proposed in the past including proportional duration adjustment [9], interpolation of models in various SRs [10-12], explicit modeling of SR effect on prosodic features [13-20,24,25], and so on.

SR is conventionally defined as words, syllables or phones per second. It is usually measured on an utterance. Since the duration of a pronunciation unit (word or syllable) is influenced by many factors, such as phonetic structure, position in sentence, speaker’s intention, SR can only be measured reliably for long utterances. But, humans can change their speaking rate as they wish in their speech. So, the utterance-based SR measure cannot always reflect the real SR of speech.

In this paper, the estimation of local inverse SR (ISR) in an utterance for Mandarin speech is addressed. Here, ISR is defined as the averaged syllable duration. The reason of using ISR is owing to its convenience to serve as a prosodic feature in TTS application. A prosodic phrase (PPh)-based SR estimation method for TTS is proposed in the study. PPh is adopted here as the segment units for ISR estimation because of their proper sizes. Syllable (SYL) and prosodic word (PW) are too short to avoid granular noise in ISR estimation, while breath group/prosodic phrase group (BG/PG) is too long to show local variations of ISR. The method is based on an existing SR-HPM prosody modeling method [16] proposed previously. The basic idea of the method is to apply the SR-HPM prosody modeling method to analyze all utterances of a large training corpus to extract their prosodic structures, and then estimate the ISRs of all prosodic phrases. The SR-HPM model is then refined using the estimated PPh ISR. A syntactic-PPh ISR model is then build and used in accompanying with the refined SR-HPM to generate prosodic features for TTS.

The contributions of this paper include: 1) A new local ISR estimation method is proposed; 2) A syntactic-local ISR model is built to determine local ISRs that are dependent on positions of prosodic units in a hierarchical prosodic structure; 3) We demonstrate the generation of speech prosody in a way of variable ISR.

The paper is organized as follows. Section 2 presents an overview of the research: a PPh-based ISR estimation method and its application to TTS. Section 3 discusses the PPh-based ISR estimation method in detail. Section 4 describes the experimental results on a large read-speech corpus. An analysis of the estimated PPh ISRs is discussed. Some conclusions are given in the last section.

2. Research Overview

Fig. 1 shows a block diagram of the proposed method of local ISR estimation and its application to the prosody generation of Mandarin TTS. The system comprises two phases: training and synthesis. In the training phase, the SR-HPM modeling method is firstly employed to build an SR-HPM model from a large training speech corpus with the utterance-based ISR being taken as an independent variable. Meanwhile, all utterances of the corpus are labeled with break and prosodic state tags. A four-layer prosodic structure for each utterance is implicitly constructed by its break tags. The prosodic structure, as shown in Fig. 2, is formed by four prosodic constituents: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath group or prosodic phrase group (BG/PG). Then, a maximum a posteriori (MAP)-based method is proposed to estimate the SRs of all PPhs of the corpus. A neural network (NN)-based syntactic-PPh ISR model is then built to describe the relation of PPh’s SRs and contextual linguistic features. Besides, the SR-HPM model is re-trained using prosodic-acoustic features normalized by the estimated PPh SRs. In the synthesis phase, the break and prosodic state tags of the input text are firstly predicted by using the re-fined SR-HPM and the given utterance-based ISR. The ISRs of all PPhs are then estimated by the syntactic-PPh ISR model. Then, SR-normalized prosodic features are generated by using the re-

10.21437/Interspeech.2018-1214

fined SR-HPM and lastly denormalized by using the estimated PPh ISRs.

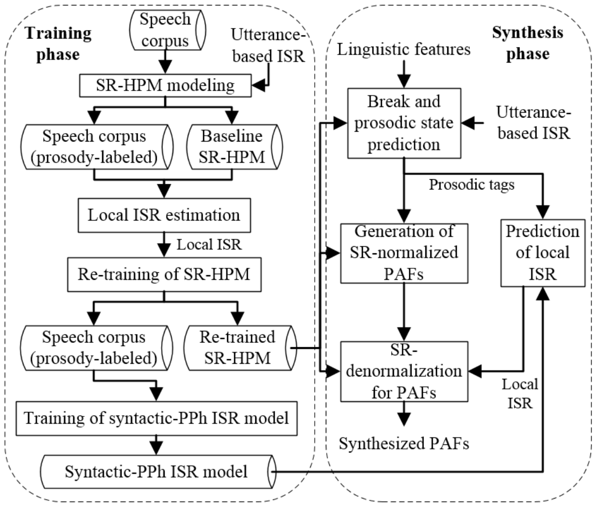

Fig. 1: A block diagram of the system using the proposed PPh ISR estimation method in the prosody generation of Mandarin TTS

In the following, we describe the proposed system in more detail. As shown in Fig.1, the system comprises five steps:

1. Training of the baseline SR-HPM and labeling the corpus
2. Estimation of local SR
3. Modeling of local SR for prosody generation in TTS
4. Re-estimation of SR-HPM
5. Prosody generation by using re-trained SR-HPM and the predicted local SR

The first step is to train the baseline SR-HPM and label all utterances with break and prosodic state tags. It first obtains the SR-normalized prosodic-acoustic features (PAFs) $\mathbf{A}'$ to suppress the effect of SR on the observed PAFs $\mathbf{A}$ by using the associated linguistic features $\mathbf{L}$, utterance-based ISRs $\mathbf{x}$, and the trained normalization functions (NFs). Then, a joint prosody labeling and modeling (PLM) algorithm is applied to simultaneously construct the SR-HPM containing five prosodic sub-models and label all utterances with the prosodic tags $\mathbf{T=\{B,P\}}$ representing the prosodic structures of utterances. The tag $\mathbf{B}$ is the break type sequence formed by seven break types {*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4} used to delimit an utterance into four types of layered prosodic constituents as shown in Figure 2: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath/prosodic phrase group (BG/PG) [20,22]. The tag set $\mathbf{P}=\{\mathbf{p},\mathbf{q},\mathbf{r}\}$ comprises three prosodic state sequences representing the states of the current syllable in higher-level prosodic constituent patterns for syllable pitch contour, syllable duration and syllable energy level, respectively [20]. Notice that the prosodic state patterns of these four prosodic constituents carry low- to high-level prosodic structure information. We therefore use prosodic states to deal with the influences of prosodic structure in our local ISR estimation.

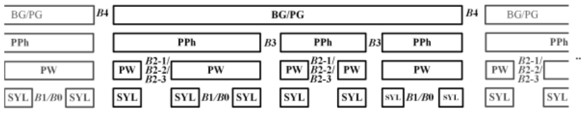

Fig. 2: Four-layer hierarchical Chinese prosodic structure [20]

The second step estimates the PPh ISRs, i.e., $\hat{\mathbf{x}}$, based on the assumption that each PPh is uttered in a constant SR which deviates closely from the SR of the intermediate upper-layer prosodic unit, i.e., BG/PG. Intuitively, we can estimate the SR of each PPh simply by averaging durations of all syllables in the PPh. However, the sample size of syllable duration will be small, in general, to result in poor ISR estimation. We therefore proposed a hierarchical MAP estimation approach to estimate local ISRs sequentially from top to lower prosodic layers to ensure that the ISR of a prosodic unit does not differ too much from the ISR of prosodic unit in its immediate upper layer. The proposed method also considers the information of tone, base syllable type and prosodic structure, which are provided by the baseline SR-HPM, to suppress the estimation bias. Specifically, we first estimate ISR of the *k*-th utterance, i.e., x_k. Then, the ISR of the *l*-th BG/PG in the *k*-th utterance, i.e., $x_{k,l}$, is estimated by the MAP method with a Gaussian prior in which the prior mean is set to be x_k. Last, the ISR of the *m*-th PPh in the *l*-th BG/PG and the *k*-th utterance, i.e., $x_{k,l,m}$, is estimated based on the prior mean $x_{k,l}$.

The third step analyzes the estimated local ISRs $\hat{\mathbf{x}}$ via exploring their patterns for different sizes of prosodic units, for different locations of BG/PG in an utterance, and for different locations of PPh within a BG/PPh, respectively. The exploration is powered by a neural net-based regression mechanism which also can serve as a local ISR predictor for the prosody generation in the TTS application.

The fourth step re-trains the SR-HPM with the estimated local ISRs $\hat{\mathbf{x}}$. Note that the NFs for suppressing SR effects on PAFs are re-trained and applied for each local prosodic unit, i.e. PPh. The sub-models are also re-trained with the local ISR.

The last step is to generate prosodic features for TTS. The break and prosodic state tags of the input text are firstly predicted by using the re-fined SR-HPM and the given utterance-based ISR. The ISR of all PPhs are then estimated by the syntactic-PPh ISR model. Then, SR-normalized prosodic features are generated by using the re-fined SR-HPM and lastly denormalized by using these predicted PPh ISRs.

3. Estimation of Local Speaking Rate

The estimation of local speaking rate is conducted in a hierarchical MAP fashion. We first refine the utterance-based ISR x_k by the hidden ISR estimation method proposed in the previous study [23]. Then, the local ISRs of the BG/PGs and PPhs are sequentially estimated as described in the second step of Section 2. Specifically, the estimation of the ISR for the *l*-th BG/PG of the *k*-th utterance is formulated by

$$x = \arg\max_{x} p(x \mid \mathbf{sd},\mathbf{t},\mathbf{s},\mathbf{B}) = \arg\max_{x} p(\mathbf{sd} \mid x,\mathbf{t},\mathbf{s},\mathbf{B})\, p(x) \quad (1)$$

where x is the local speaking rate to be estimated for the *l*-th BG/PG of the *k*-th utterance; $\mathbf{sd}=\{sd_n\}_{n=1\sim N}$, $\mathbf{t}=\{t_n\}_{n=1\sim N}$, $\mathbf{s}=\{s_n\}_{n=1\sim N}$ and $\mathbf{B}=\{B_n\}_{n=1\sim N}$ are sequences of syllable duration, tone, base syllable type and break type, respectively; n is the syllable index; N is the number of syllables of the BG/PG. The probability $p(\mathbf{sd} \mid x,\mathbf{t},\mathbf{s},\mathbf{B})$ is the likelihood function describing the distribution of $\mathbf{sd}$ given information of x, $\mathbf{t}$, $\mathbf{s}$, and prosodic structure represented by the break type sequence $\mathbf{B}$. The prior probability $p(x)$ is modeled by a normal distribution, i.e., $x \sim N(x_k, v_{x_k})$. In the likelihood function, we consider several affecting factors that influence the variation of syllable duration, including tone, base-syllable type, SR, and prosodic state. Following the definition in our previous studies, prosodic state is conceptually defined as the state in a prosodic phrase and accounts for prosodic variation resulted from the prosodic structure which is represented by the break type sequence. Then, the affecting patterns (APs) associated with the above-mentioned affecting factors are defined to control the increase or decrease of syllable duration. Based on the assumption that these APs are combined additively, the syllable duration is expressed by

$$sd_n = sd'_n + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + x \qquad (2)$$

where γ_{t_n} , γ_{s_n} , γ_{q_n} are APs of tone (t_n), base syllable type (s_n) and prosodic state (q_n); and sd'_n is the residual modeled by a zero-mean normal distribution. Notice that the prosodic state is absent in the likelihood function and hence treated as a latent variable which is dependent on the break type sequence. We therefore introduce the expectation-maximization (EM) algorithm to solve the problem in Eq. (1) based on the MAP criterion, i.e.,

$$x = \arg\max_{x} \sum\nolimits_{\mathbf{q}} p(\mathbf{q} \mid \mathbf{sd}, x', \mathbf{t}, \mathbf{s}, \mathbf{B}) \ln\left[p(\mathbf{sd} \mid \mathbf{q}, x, \mathbf{t}, \mathbf{s}, \mathbf{B})\, p(x) \right] \qquad (3)$$

where $p(\mathbf{q} \mid \mathbf{sd}, x', \mathbf{t}, \mathbf{s}, \mathbf{B})$ is the a posterior probability of the prosodic state; x' is the old estimate of ISR; $p(\mathbf{sd} \mid \mathbf{q}, x, \mathbf{t}, \mathbf{s}, \mathbf{B})$ is the new likelihood function which is elaborated with the additive property shown in Eq. (3):

$$p(\mathbf{sd} \mid \mathbf{q}, x, \mathbf{t}, \mathbf{s}, \mathbf{B}) \approx p(\mathbf{sd} \mid \mathbf{q}, x, \mathbf{t}, \mathbf{s}) = \prod\nolimits_{n=1}^{N} p(sd_n \mid q_n, x, t_n, s_n)$$
$$= \prod\nolimits_{n=1}^{N} N(sd_n \mid \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n} + x, v) \qquad (4)$$

For simplicity, the posterior probability is approximated by assuming that the prosodic state depends only on the break type sequence which specifies the prosodic structure, i.e.

$$p(\mathbf{q} \mid \mathbf{sd}, x', \mathbf{t}, \mathbf{s}, \mathbf{B}) \approx p(\mathbf{q} \mid \mathbf{B}) = \prod\nolimits_{n=1}^{N} p(q_n \mid \mathbf{B}) \qquad (5)$$

The probability $p(q_n \mid \mathbf{B})$ can be estimated by the forward-backward calculation with the probability $p(q_n \mid q_{n-1}, B_{n-1}, B_n)$. Note that the APs of γ_{t_n} , γ_{s_n} and γ_{q_n} , and the probability $p(q_n \mid q_{n-1}, B_{n-1}, B_n)$ can be obtained after the baseline SR-HPM is trained in the second step with break and prosodic state sequences being labeled.

4. Analysis on Local Speaking Rate

4.1. Experimental Database

The database for examining the proposed ISR estimation is the same one as used in our previous study [16] - a female Mandarin read-speech corpus comprising four parallel sub-corpora of fast, normal, medium, and slow SRs. The texts of all utterances in these four parallel sub-corpora are short paragraphs which are excerpted from news and articles. The maximum and minimum lengths of these utterances are 270 and 80 syllables, and the average length is 138 syllables. The database is divided into a training set with 183,795 syllables for SR-HPM training and a test set with 19,951 syllables for prosody generation experiment.

4.2. Analysis by Synthesis

An analysis-by-synthesis method is adopted here to explore the local ISR patterns of BG/PGs and PPhs. First, we analyze the patterns of BG/PGs in an utterance, i.e., $x_{k,l}$, by a neural net-based regression mechanism. The neural net is in a structure of one hidden layer with hyperbolic tangent activation and the output layer with one node that represents the local ISR of BG/PG ($x_{k,l}$). The input feature vector is composed of the utterance-based ISR (ISR_Utt), number of syllables in the utterance (#S_Utt), number of BG/PGs in the utterance (#B_Utt), number of syllables in the current BG/PG (#S_B), and the normalized BG/PG forward position index (Pos_B) defined as (l-1)/(L-1), where L represents the number of BG/PGs in the utterance.

Then, we perform the neural net-based regression to analyze the patterns of PPh ISRs in an utterance. The input features used in the neural net include the ones used in the neural net for analyzing the ISR patterns of BG/PG, the number of PPhs in the current BG/PG (#P_B), the number of syllables in the current PPh (#S_P), and the normalized PPh forward position index (Pos_P).

Table 1 shows the average total residual errors (TREs) resulted from using various input feature combinations. For BG/PG ISR patterns, the average TREs for the feature combinations comprising the lengths of utterance and/or BG/PG are generally lower than the ones without length features. The TREs for PPh are better than those of BG/PG. Table 2 shows the statistics of the lengths for various prosodic constituents. As shown in the table that the average number of syllables in a PPh (BG/PG) is 9.4 (22.3) for fast speech, and decreases to 7.5 (13.9) for slow speech.

Table 1: Average total residual errors

	ISR_Utt #B_Utt Pos_B	**#S_Utt**	**#S_B**	**#P_B Pos_P**	**#S_P**	**TREs Training /Test**
	v					1.09/1.20
BG/PG	v	v				1.12/1.24
NN	v		v			1.14/1.18
	v	v	v			1.02/1.14
PPh	v	v	v	v		0.93/0.98
NN	v	v	v	v	v	0.89/0.94

Table 2: Statistics of the lengths of various prosodic units

Prosodic unit	**Length unit**	**fast**	**normal**	**medium**	**slow**
utterance	SYL	137.7	137.9	138.0	137.7
	BG/PG	6.2	7.1	7.1	9.9
	PPh	14.8	15.8	15.9	18.4
BG/PG	SYL	22.3	19.5	19.4	13.9
	PPh	2.4	2.2	2.2	1.9
PPh	SYL	9.4	8.8	8.7	7.5

Fig. 3 shows typical patterns of BG/PG and PPh ISR estimates for an utterance spoken in four representative utterance-based ISRs of 0.173 (fast), 0.186 (normal), 0.207 (medium) and 0.217 (slow) seconds/syllable. It can be found from the figure that the BG/PG-based ISRs deviate closely around the utterance-based ISR. For each BG/PG-based ISR, PPh-based ISRs distribute closely around it in fast to slow patterns for most cases.

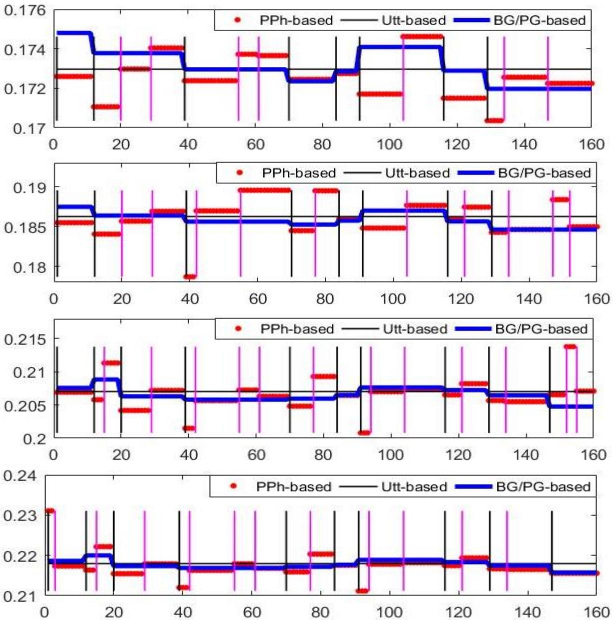

Fig. 3: Typical local ISR patterns for an utterance spoken in utterance-based ISRs of 0.173 (fast), 0.186 (normal), 0.207 (medium), and 0.217 (slow) seconds/syllable.

5. Application to Prosody Generation

We conduct several prosody generation experiments to verify that the estimated local ISRs are meaningful and could accurately describe the speaker’s speaking rate variations. We design two experiments: an oracle one and a real one. The oracle prosody generation experiment produces the synthesized PAFs given with the correct break tags labeled by the trained SR-HPM and the estimated local ISR by the method presented in Section 3. The real prosody generation experiment is conducted in a real TTS application scenario to produce the PAFs by using the refined SR-HPM and the local ISR predictor given with the correct utterance-based ISR.

The purpose of the oracle experiment is to examine if the estimated local ISR could accurately model the prosodic variations in terms of the objective measures. The objective measures used here are the root-mean-square error (RMSE) and the correlation coefficients calculated with the true and generated PAFs. We compare the performances of the utterance-based, BG/PG-based and the PPh-based ISR estimations, and the associated estimation methods: RAW, EM, and EM-MAP. The RAW method is to simply estimate the ISR by averaging syllable durations of a prosodic unit. The EM-MAP method estimates the local ISR by Eq. (3), while the EM method estimates the local ISR by Eq. (3) without the prior probability $p(x)$. It can be seen from Table 3(a) that the proposed PPh-based ISR estimation with EM-MAP yielded the lowest RMSEs for *sd* and **sp**, and the highest correlation coefficient for *sd*. The performances for *se* and *pd* are degraded in the cases of the BG/PG-based and PPh-based estimations. In general, the lowest RMSE and the highest correlation coefficient are achieved by the EM-MAP estimation method, followed by the EM and RAW estimation methods.

Table 3(b) shows the RMSE and correlation coefficient between the PAFs generated by the real prosody generation experiments and the true PAFs. We compare the results by the three configurations of real pro: UTT-based RAW, UTT-based EM, and PPh-based EM-MAP. The results by the UTT-based RAW configuration are obtained by the PAFs generated by the baseline SR-HPM with linguistic features and utterance-based raw ISR. The UTT-based EM results are obtained by the PAFs generates by the re-trained SR-HPM with linguistic features and the utterance-based EM-estimated ISR. The PPh-based EM-MAP results are obtained by the re-trained SR-HPM with linguistic features and PPh ISR predicted by the local ISR predictor. As shown in the table that PPh-based EM-MAP has the best performance.

An informal listening test confirmed that the synthesized speech of the new method using PPh-based ISR estimates is more vivid than that of the existing SR-HPM method using a given utterance-based ISR.

Table 3: RMSEs and correlation coefficients between the predicted and true PAFs under the conditions of (a) with **correct** break and **correct** local ISR, and (b) with **predicted** break and **predicted** local ISR.

(a)		UTT-based[a]		BG/PG-based[b]			PPh-based[c]		
		RAW[d]	EM[e]	RAW	EM	EM-MAP[f]	RAW	EM	EM-MAP
RMSE	*sd*[g]	48.2	47.7	47.7	48.3	47.2	48.0	46.2	45.4
	sp[n]	.1472	.1467	.1650	.1469	.1469	.1472	.1465	.1463
	se[i]	3.54	3.53	3.56	3.56	3.52	3.57	3.55	3.56
	pd[j]	55.2	55.2	58.2	56.8	55.5	61.9	60.6	59.6
COR[k]	*sd*	.779	.784	.783	.784	.790	.786	.802	.810
	sp[l]	.776 .815 .631 .524	.776 .814 .631 .524	.775 .815 .634 .524	.774 .815 .631 .524	.776 .816 .632 .527	.773 .815 .633 .526	.780 .815 .633 .525	.779 .816 .632 .527
	se	.887	.888	.887	.887	.890	.887	.887	.887
	pd	.954	.954	.948	.951	.954	.941	.943	.945

(b)	RMSE				COR			
	sd	**sp**	*se*	*pd*	*sd*	**sp**[m]	*se*	*pd*
UTT-based RAW	49.1	.1597	3.63	88.2	.770	[.727 .774 .600 .494]	.881	.881
UTT-based EM	48.8	.1580	3.63	87.4	.773	[.731 .773 .602 .501]	.882	.881
PPh-based EM-MAP	48.0	.1578	3.63	87.6	.783	[.734 .775 .602 .498]	.883	.880

[a] UTT-based: SR-HPM with utterance-based hidden ISR.

[b] BG/PG-based: SR-HPM trained with BG/PG-based ISR.

[c] PPh-based: SR-HPM trained with PPh-based ISR.

[d] RAW: Raw ISR obtained by simply averaging syllable duration.

[e] EM: ISR estimated with EM algorithm.

[f] EM-MAP: ISR estimated by the EM algorithm with MAP criterion.

[g] sd: second, [h] sp: logHz, [i] se: dB, [j] pd: second

[k] COR: correlation coefficient

[l] sp: CORs of four-dimensional logF0 contour

6. Conclusions

A new local speaking rate method has been discussed in this paper. It is based on the four-layer prosodic structure proposed previously to explore the local speaking rate variations on high-level prosodic constituents of BG/PG and PPh. Experimental results showed that the PPh-based speaking rate estimates distribute closely around the utterance-based speaking rate. As applying the proposed method to Mandarin TTS, more vivid synthesized speech can be obtained.

7. Acknowledgements

This work was supported by the MOST of Taiwan under Contract No. MOST-106-2221-E-305-010.

8. References

[1] Jacewicz, Ewa, Robert A. Fox, Caitlin O'Neill, and Joseph Salmons. “Articulation Rate across Dialect, Age, and Gender.” Language Variation and Change 21, no. 2 (2009): 233–56

[2] H. Fujimura, T. Masuko, and M. Tachimori, “A duration modeling technique with incremental speech rate normalization,” in *Proc. INTERSPEECH’10*, Makuhari, Japan, Sep. 2010, pp. 2962–2965.

[3] T. Pfau, R. Faltlhauser, and G. Ruske, “A combination of speaker normalization and speech rate normalization for automatic speech recognition,” in *Proc. ICSLP’00*, Beijing, China, Oct. 2000, pp. 362–365.

[4] S. M. Chu and D. Povey, “Speaking rate adaptation using continuous frame rate normalization,” in *Proc. ICASSP’10*, Dallas, TX, USA, Mar. 2010, pp. 4306–4309.

[5] D. Jouvet, D. Fohr, and I. Illina, “About handling boundary uncertainty in a speaking rate dependent modeling approach,” in *Proc. INTER-SPEECH’11*, Florence, Italy, Aug. 2011, pp. 2593–2596.

[6] T. Shinozaki and S. Furui, “Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation,” in *Proc. ASRU’03*, Thomas, U.S. Virgin Islands, Nov. 2003, pp. 417–422.

[7] 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.

[8] R. Lotfian, C. Busso, “Emotion recognition using synthetic speech as neutral reference,” in *Proc. ICASSP’15*, Brisbane, QLD, Australia, Apr. 2015, pp. 4759–4763.

[9] 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*, Florence, Italy, Aug. 2011, pp. 1845–1848.

[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*, Beijing, China, Aug. 2009, pp. 148–153.

[11] K. Iwano, M. Yamada, T. Togawa, and S. Furui, “Speech-rate variable HMM-based Japanese TTS system,” in *Proc. TTS’02*, Santa Monica, CA, USA, Sep. 2002.

[12] 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*, Makuhari, Japan, Sep. 2010, pp. 2186–2189.

[13] 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.

[14] 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

[15] 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.

[16] S. H. Chen, C. H. Hsieh, C. Y. Chiang, H. C. Hsiao, Y. R. Wang, and Y. F. Liao, H. M. Yu, “Modeling of Speaking Rate Influences on Mandarin Speech Prosody and Its Application to Speaking Rate-controlled TTS,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1158–1171, May. 2014.

[17] I. B. Liao, C. Y. Chiang, Y. R. Wang, S. H. Chen, “Speaker Adaptation of SR-HPM for Speaking Rate-Controlled Mandarin TTS,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2046–2058, Aug. 2016.

[18] 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*, Singapore, Sept. 2014, pp. 511-515.

[19] I. B. Liao, C. Y. Chiang, and S. H. Chen, “Structure maximum a posteriori speaker adaptation of speaking rate-dependent hierarchical prosody model for Mandarin TTS,” in *Proc. ICASSP’16*, Shanghai, China, Mar. 2016.

[20] C. Y. Chiang, "Cross-Dialect Adaptation Framework for Constructing Prosodic Models for Chinese Dialect Text-to-Speech Systems," I*EEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 108-121, Jan. 2018.

[21] 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, 1990.

[22] 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.

[23] G. T. Liou, C. Y. Chiang, Y. R. Wang, and S. H. Chen, “Estimation of Hidden Speaking Rate,” accepted by *Speech Prosody*, Jun. 2018.

[24] H. Quené, “Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo,” *J. Acoust. Soc. Amer.*, vol. 123, no. 2, pp. 1104–1113, Feb. 2008.

[25] J. Trouvain, *Tempo variation in speech production: Implications for speech synthesis*, PhD thesis, University of Saarland, 2003.