9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Estimation of Hidden Speaking Rate

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

Hidden speaking rate is proposed in this paper. In contrast to traditional raw speaking rate estimation that simply averages number of syllable or phone per second with or without pauses, the proposed hidden speaking rate is estimated by normalizing effects of lexical information and prosodic structure based on the existing speaking rate-dependent hierarchical prosodic model (SR-HPM). The significance of the proposed hidden speaking rate is exemplified by analysis on the speaking rate estimation for a Mandarin speech database containing four parallel speech corpora of a female professional announcer with fast, normal, medium and slow speaking rates. By conducting prosody generation experiment on the same speech corpus, the hidden speaking rate is proved to be more meaningful and accurate to represent speaker's intended or underlying speaking rate than conventional raw speaking rate.

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

1. Introduction

Speaking Rate (SR), a prosody feature, influences many speech phenomena such as pause duration, syllable duration, pitch contour shape, and so on. Conventionally, speech rate and articulation rate (AR) are two SR measures defined as the number of output units per unit of time. The difference between speech rate and AR is that speech rate includes pause intervals while AR does not. AR determines the pace at which speech segments are actually produced and does not take into account speaker-specific ways of conveying information, such as hesitations, pausing, emotional expressions, and so on [1]. On the other hand, speech rate contains more global speaker characteristics including frequency of pausing.

Estimation and modeling SR are useful for many speech applications, such as automatic speech recognition (ASR), emotion recognition, and text-to-speech system (TTS). For ASR, acoustic modeling (AM) is less robust for very fast or slow speech. Many methods have been proposed to enhance the robustness of AM, including utilization of durational information [2], SR-normalization of spectral features [3,4], modeling of pronunciation variations [5], adaptation of HMM's mixture weights and transition probabilities [6], use of parallel AMs of various SRs [7]. [8] used estimated SR to assist in emotion recognition. For TTS, generating speech in user-defined or controllable SR makes the synthesized speech more vivid and suitable for the various application, e.g., fast speech for visually-impaired people and slow speech for language learners. Existing methods for modeling SR in TTS are proportional duration adjustment [9], interpolation of models in various SRs [10-12], and explicit modeling of SR effect on prosodic features [13-20].

It is found that the mentioned-above previous studies in SR modeling measured speaking rate by using simply averaging number of syllables or phones in an observed duration with or without pause duration. Some studies [14-20] also measure SR by averaging syllable duration per second, i.e., inverse SR (ISR) to facilitate duration modeling. However, observed (or raw) syllable or phone durations are influenced by lexical contents and prosodic structures of utterances. This indicates that SR estimated by averaging raw duration would be biased by lexical and prosodic structure information. The true SR would be view as a hidden variable because the only observations are durations of syllables or phones.

In this paper, we propose to estimate the so-called hidden inverse SR (hidden ISR) based on the existing speaking rate-dependent hierarchical prosodic model (SR-HPM). The main idea is that we consider SR is influenced by factors of lexical information of tone and base syllable type and prosodic structure. The hidden ISR is estimated by syllable durations being normalized by the above factors. The role of the SR-HPM is used to obtain the prosodic structure of input utterances and the patterns of the affecting factors. The significance of the proposed hidden ISR is exemplified, and prosody generation experiment is conducted to examine the usefulness of the proposed hidden speaking rate.

This paper is organized as follows. Section 2 describes the overall research process of this study. Section 3 presents the estimation of the proposed hidden ISR. Examination of the proposed hidden ISR is stated in Section 4. Some conclusions and future studies are presented in the final section.

2. Overall Research Process

Fig. 1 shows the proposed estimation of hidden ISR and the applications to prosody modeling and generation.

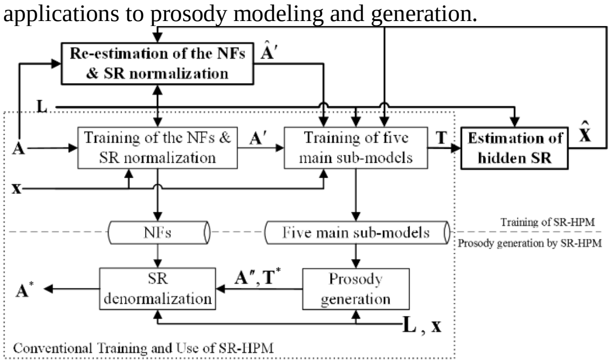
Re-estimation of the NFs & SR normalization
$\hat{A}'$
L
A
Training of the NFs & SR normalization
A′
Training of five main sub-models
T
Estimation of hidden SR
$\hat{X}$
x
NFs
Five main sub-models
Training of SR-HPM
Prosody generation by SR-HPM
A*
SR denormalization
A″, T*
Prosody generation
L , x
Conventional Training and Use of SR-HPM

Fig. 1: *The proposed estimation of hidden ISR and the applications to prosody modeling and prosody generation.*

10.21437/SpeechProsody.2018-120

The research is conducted in five steps. We describe them as follows. The first two steps are conventional training procedure of SR-HPM. The step 1 constructs normalization functions (NFs) to suppress the effect of SR on observed prosodic-acoustic features (PAFs), i.e., **A**, given with the ISR, i.e., **x**, and the associated linguistic features (**L**). Here, the ISR is defined as an average syllable duration calculated for an utterance (unit: second per syllable, s/syl), so-called ***raw ISR***. The NFs are constructed for four types of PAFs: syllable pitch contour (**sp**), syllable duration (*sd*), syllable energy level (*se*) and inter-syllable pause duration (*pd*). Each **sp** is represented by a four-dimensional vector [21]. The NF for the **sp** vector is a tone-dependent and dimension-dependent z-normalization function with mean and variance given by two first-order polynomials of **x**. The NF for *sd* is a z-normalization function with mean and variance given by two second-order polynomials of **x**. The NF for pause duration is a Gamma distribution normalization function, with its two parameters calculated from mean and variance represented by two second-order polynomials of **x**. The NF for *se* is a z-normalization function with utterance-dependent mean and variance that does not consider the SR effect.

The step 2 conducts a joint prosody labeling and modeling (PLM) algorithm to obtain five main prosodic sub-models and prosodic tags that represent prosodic structures of input utterances, i.e., **T = {B, P}**. The tag **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 four types of layered prosodic constituents: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath/prosodic phrase group (BG/PG) [20,22]. The tag set **P** = {**p**, **q**, **r**} comprises three prosodic state sequences representing the states of the current syllable in higher-level prosodic constituent patterns for syllable pitch contour (**p**), syllable duration (**q**) and syllable energy level (**r**), respectively [20]. The inputs of the training for the five sub-models are SR-normalized PAFs ($\mathbf{A}'$), ISR (**x**) and linguistic feature (**L**). The trained NFs and the trained five sub-models can be used in prosody generation. The prosody generated by this step is taken as the baseline for comparison.

The step 3 estimates the hidden ISR ($\hat{\mathbf{x}}$) given with **A**, **L**, and information about tone, base syllable type and a prosodic structure represented by the break type tag (**B**) which is obtained by the conventional SR-HPM training procedure (step 2). The step 4 re-estimates the *sd* NFs with the estimated hidden ISR by considering the same factors used for estimating the hidden ISR. All the other NFs are trained in the conventional method but with the hidden ISR, $\hat{\mathbf{x}}$. The step 5 re-trains the five sub-models (with the hidden ISR, $\hat{\mathbf{x}}$) and these re-trained sub-models and the re-estimated NFs are used to conduct prosody generation experiments. The prosody generated in this step is used to examine the usefulness of the proposed hidden ISR.

3. The Proposed ISR

3.1. Estimation of Hidden ISR

Recall that the *sd* of each syllable in SR-HPM is modeled by an additive model with the *sd* NF:

$$sd_n = (sd'_n + \gamma_{t_n} + \gamma_{s_n} + \gamma_{q_n})\sigma(x) / \sigma_g + x \qquad (1)$$

where γ_{t_n} , γ_{s_n} , γ_{q_n} represent affecting patterns (AP) of tone (t_n), base syllable type (s_n) and prosodic state (q_n); sd'_n is the modeling residual; $\sigma(x)$ is the *sd* NF expressed by

$$\sigma(x) = ax^2 + bx + c \qquad (2)$$

; σ_g is the global *sd* standard deviation. By (1), we may directly obtain the hidden ISR by moving *x* to the left-hand side of (1) and moving the rests to the right-hand side. However, the labeled prosodic state (q_n) for each syllable by the SR-HPM tends to be quantized value for the residual of the following value:

$$(sd_n - x)\sigma_g / \sigma(x) - \gamma_{t_n} - \gamma_{s_n} \qquad (3)$$

This makes sd'_n zero mean and very small in variance. Therefore, sd_n is well-fitted by γ_{t_n} , γ_{s_n} , γ_{q_n} , *x*, $\sigma(x)$ and σ_g . In this situation, the prosodic state sequence **q** and its affecting pattern γ_{q_n} overfit the observed *sd*. Directly obtaining the hidden ISR by moving *x* to the left-hand side of (2) will not refine the ISR. To overcome this overfitting problem, we estimate the AP of the prosodic state by information given with the labeled break sequence **B**. In other words; we estimate a more robust prosodic state pattern by the given prosodic structure represented by break type sequence. Specifically, the hidden ISR is estimated by:

$$x = \tfrac{1}{N}\sum\nolimits_{n-1}^{N}\{sd_n - \gamma_{t_n} - \gamma_{s_n} - \sum\nolimits_{q_n}[p(q_n \mid \mathbf{B})\gamma_{q_n}]\} \qquad (4)$$

where $p(q_n \mid \mathbf{B})$ is the posterior probability of duration prosodic state q_n conditioned on the prosodic structure represented by break sequence **B;** *N* is number of syllable of an utterance. The probability $p(q_n \mid \mathbf{B})$ can be estimated by forward/backward calculation with probabilities $p(q_n \mid q_{n-1}, B_{n-1}, B_n)$. 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 SR-HPM is trained in the step 2 with break and prosodic state sequences labeled.

3.2. Re-estimation of the Syllable Duration NF

The parameters of conventional *sd* NF, i.e. *a*, *b* and *c* in (2), are obtained by fitting points of (x_k, σ_k) under a minimum mean square error criterion; where x_k and σ_k are the ISR and *sd* standard deviation of *k*-th utterance. In this study, the parameters of the *sd* NF are estimated by considering factors of tone (**t**), base syllable (**s**) and break type (**B**):

$$\begin{aligned}\{a,b,c\}^* &= \arg\max_{a,b,c} p(a,b,c \mid \hat{\mathbf{x}}, \boldsymbol{\sigma}, \mathbf{t}, \mathbf{s}, \mathbf{B}) \\ &= \arg\max_{a,b,c} p(\boldsymbol{\sigma} \mid a,b,c,\hat{\mathbf{x}}, \boldsymbol{\sigma}, \mathbf{t}, \mathbf{s}, \mathbf{B})\, p(a,b,c) \\ &= \arg\max_{a,b,c} \prod\nolimits_{k=1}^{K} p(\sigma_k \mid a,b,c,\hat{x}_k, \mathbf{t}_k, \mathbf{s}_k, \mathbf{B}_k)\, p(a,b,c)\end{aligned} \qquad (5)$$

where *K* is the number of training utterances; $p(a,b,c)$ is the posterior probability; $\hat{x}_k$ is hidden ISR of the *k*-th utterance. By (1), we may assume

$$\sigma_k \sim N(std((sd'_{n,k} + \gamma_{t_{n,k}} + \gamma_{s_{n,k}} + \gamma_{q_{n,k}})\sigma(x_k) / \sigma_g \mid \mathbf{B}_k), v) \qquad (6)$$

Suppose $sd'_{n,k}$, $\gamma_{t_{n,k}}$, $\gamma_{s_{n,k}}$ and $\gamma_{q_{n,k}}$ are orthogonal to each other, and only prosodic state is dependent on $\mathbf{B}_k$, then we have

$$\begin{aligned}&Var(sd'_{n,k} + \gamma_{t_{n,k}} + \gamma_{s_{n,k}} + \gamma_{q_{n,k}} \mid \mathbf{B}_k) \\ &= Var(sd'_{n,k}) + Var(\gamma_{t_{n,k}}) + Var(\gamma_{s_{n,k}}) + Var(\gamma_{q_{n,k}} \mid \mathbf{B}_k)\end{aligned} \qquad (7)$$

The variances of $sd'_{n,k}$, $\gamma_{t_{n,k}}$ and $\gamma_{s_{n,k}}$ can be obtained directly by tone and base syllable APs and statistics of utterances' tone and base syllable sequences. The variance $Var(\gamma_{q_{n,k}} | \mathbf{B}_k)$ can be estimated by the posterior probability of duration prosodic state q_n conditioned on the prosodic structure represented by break sequence **B**.

4. Experimental Results

4.1. Experimental Database

The database for examining the proposed ISR is the same one as used in our previous study [16] - a female Mandarin speech corpus that contains four parallel sub-corpora of fast, normal, medium, and slow SRs. The four parallel sub-corpora were recorded with the short text paragraphs which were excerpted from news and articles. The maximum and minimum lengths of the 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 of Estimated Hidden ISR

We first examine the estimated hidden ISRs for each of the sub-corpora. Fig. 2 shows histograms of raw ISRs and the hidden ISRs. It can be seen from the figure that each of the four speech sub-corpora has narrower distribution in the hidden ISR than the one in the raw ISR. Table 1 shows average raw ISRs and hidden ISRs for the four sub-corpora. We found that the estimated hidden ISRs are slower than the raw ISRs for fast and normal speech corpora while they are faster for medium and slow speech corpora. Table 2 shows variances of raw ISRs and the proposed hidden ISRs for each of the four sub-corpora. It is found that the variances of hidden ISR are smaller than the ones of raw ISR for each of the corpora. This result is in accordance with the histograms of Fig. 1. Table 2 also shows the variances of the hidden ISRs estimated with each single factors of tone, base syllable type and prosodic structure, and combinations of tone and base syllable type factors. We may conclude from the above observations that the proposed hidden ISR may be more suitable to represent speaker's underlying SR than raw ISR because variances of hidden ISR for speech corpus of each speaking rate range (fast, normal, medium, and slow) is lower than ones of raw ISRs. Besides, the proposed method can suppress the effect of tone, base syllable type and prosodic structure on ISR estimation. This suppression reduced very fast and very slow estimated ISR for the fast and slow speech corpora, which may be over- or under-estimated by the raw ISR estimation.

Then, we analyze how the factors of tone and prosodic structure affect the estimation of hidden ISR. First, we analyze the correlation coefficient between the occurrence of tone in an utterance and ISR difference between the proposed hidden ISR and raw ISR. This correlation coefficient is defined as:

$$\rho_{P_t,\Delta X} = \mathrm{cov}(p_{t,k}, \Delta x_k) / (\sigma_{P_t} \sigma_{\Delta X}) \quad (8)$$

$$\Delta x_k = \hat{x}_k - x_k \quad (9)$$

where $p_{t,k}$ is the probability (frequency) of tone t in k-th utterance; $\hat{x}_k$ and x_k are respectively the proposed hidden ISR and raw ISR of k-th utterance. Table 3 shows the correlation coefficients $\rho_{P_t,\Delta X}$ for t=1~5. It is found that frequencies of tones 3 and 5 positively correlated with the ISR difference Δx_k while those of tones 1 and 2 are negatively correlated. This result reflects the fact that tones 3 and 5 intrinsically have short syllable duration while tones 1 and 2 have long ones. These intrinsic long or short syllable durations would affect the accuracy of ISR estimation as we do not remove their effect on averaging syllable duration. For example, it an utterance contains more tone-5 syllables, the raw ISR will be under-estimated as a faster ISR.

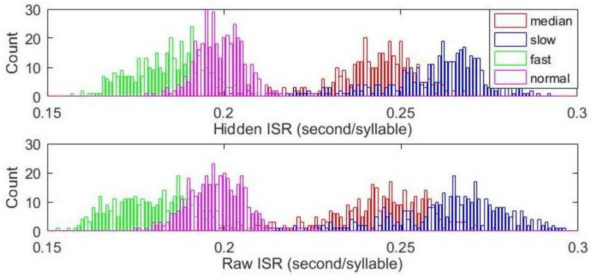

Fig. 2: *Histograms of raw ISRs and the hidden ISRs.*

Table 1: *Average raw and hidden ISRs for the four speech corpora*

	fast	normal	medium	slow
Raw ISR	0.180	0.198	0.264	0.244
Hidden ISR	0.184	0.199	0.260	0.242

Table 2: *Variances of raw ISRs and the proposed hidden ISRs for each of the four sub-corpora. The hidden ISR estimated with all factor are denoted as +t+s+B while hidden ISRs estimated with single factors of tone, base syllable type and prosodic structure are respectively denoted as +t, +s and +B. The combined factors of tone and base syllable type is denoted as +t+s. The numbers in brackets are the ratio of the variance of hidden ISR to the variance of raw ISR.*

		fast	normal	medium	slow
Raw ISR		1.271	0.629	1.822	2.584
Hidden ISR with various factors	+t+s+B	0.938 (.74)	0.425 (.67)	1.423 (.78)	2.138 (.82)
	+t	1.099 (.86)	0.536 (.85)	1.660 (.91)	2.435 (.94)
	+s	1.163 (.91)	0.495 (.79)	1.749 (.96)	2.498 (.96)
	+B	1.113 (.88)	0.587 (.93)	1.596 (.88)	2.315 (.90)
	+t+s	1.038 (.82)	0.448 (.71)	1.634 (.90)	2.396 (.93)

Table 3: *The correlation coefficient ($\rho_{P_t,\Delta X}$) between tone frequency and ISR difference Δx_k between hidden ISR and raw ISR.*

Tone *t*	1	2	3	4	5
$\rho_{P_t,\Delta X}$	-0.27	-0.57	0.30	0.05	0.90

Next, we analyze effect of prosodic structure on ISR estimation. To simplify the analysis, probability of break type b in k-th utterance, i.e. $p_{b,k}$, is used to measure correlation with the ISR difference Δx_k. Table 4 shows the correlation between $p_{b,k}$ and Δx_k, i.e. $\rho_{P_b,\Delta X}$ for each break type.

Table 4: *The correlation coefficient ($\rho_{P_b,\Delta X}$) between break frequency and ISR difference Δx_k between hidden ISR and raw ISR.*

Break *b*	B0	B1	B2-1	B2-2	B2-3	B3	B4
$\rho_{P_b,\Delta X}$	0.46	-0.02	0.25	-0.24	0.03	-0.54	0.05

It is found that higher occurrence of *B*3 and *B*2-2 would result in faster ISR estimated by the proposed method while more *B*0 in an utterance result in slower hidden ISR. Recall that *B*3 and *B*2-2 are break types with median and short pauses. These two break types are also signaled by pre-boundary syllable duration lengthening. The break type *B*0 is defined as a tightly-coupled intra- prosodic word syllable juncture and is following a shorter syllable within a prosodic word. Note that occurrences of break types are dependent on syntax and semantics of utterance's content. The result shown here indicates that the proposed estimation of hidden ISR can remove part of effect resulted from utterance's prosodic structure or text content.

4.3. Analysis of Re-estimated sd NF

Fig. 3 shows the *sd* NF estimated by the conventional method and the proposed method. The re-estimated *sd* NF has the same trend with the original NF but has a little larger value. This figure also shows the observed utterance-wise syllable duration standard deviation, i.e. σ_k , and the normalized σ_k , i.e.

$$\hat{\sigma}_k = \sigma_k \sigma_g / \sqrt{Var(sd_{n,k}) + Var(\gamma_{t_{n,k}}) + Var(\gamma_{s_{n,k}}) + Var(\gamma_{q_{n,k}})} \quad (10)$$

The scatter plot of $\hat{\sigma}_k$ is shown in Fig. 3. Note that the re-estimation of *sd* NF by (5) is analogous to a curve fitting problem that the polynomial function $\sigma(x_k)$ matches the points of $(x_k, \hat{\sigma}_k)$. It is found that $\hat{\sigma}_k$'s have smaller variance than σ_k 's have around the estimated NFs. These smaller variances are because of the consideration of affecting factors of tone, base syllable type and prosodic structure in the equation (5). The considered factors would affect the estimation of ISR. By normalizing the effect of the factor, the re-estimated *sd* NF could be more robust.

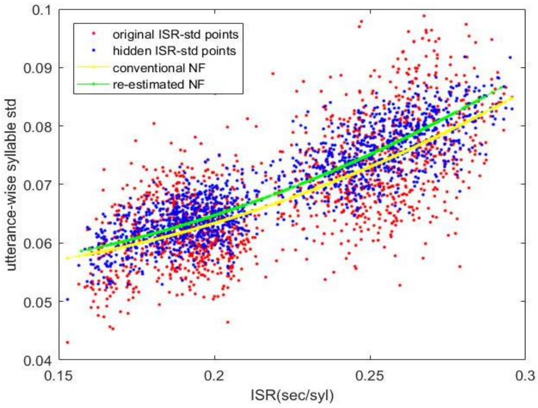

Fig. 3: *The conventional and re-estimated sd NFs.*

4.4. Analysis of Modelling Error and Break Labeling

Table 5 shows the mean squared errors (MSEs) of reconstructed pitch contour (**sp**), syllable duration (*sd*) and energy level (*se*). These MSEs are calculated by the difference between observed PAFs and the reconstructed PAFs by parameters of the five sub-models and denormalization by the NFs. It can be seen from the table that **sp**, *sd,* and *se* are all very small compared with the variance of the modeled prosodic features. The RMSEs by the re-trained SR-HPM are very close to the one by the original SR-HPM. This is because most of the variances of the modeled data have been reduced by APs of prosodic states.

Table 6 displays the correlation matrix of break types labeled by the re-trained SR-HPM with the hidden ISR and the original SR-HPM with raw ISR. It is found that break types interchange more frequently within pause-related break class (i.e., *B*2-2, *B*3 and *B*4) and non-pause-related break class (i.e., *B*0, *B*1, *B*2-2, and *B*2-3). This result indicates that the ISR difference between hidden ISR and raw ISR could affect the labeling of prosodic structure. Table 7 shows the RMSEs of reconstructed pause durations (*pd*). The reconstructed *pd* by the proposed method has smaller RMSEs than the ones by the conventional method in all break type cases. We may conclude that the proposed hidden ISR is a more accurate estimation of ISR that affect the distribution of *pd*.

Table 5: *MSEs of the reconstructed prosodic-acoustic features of syllable pitch contour, syllable duration, and syllable energy level.*

	sp $(\times 10^{-4}(\log - Hz)^2)$	*sd* (ms^2)	*se* (dB^2)
Variance	564	6000	36.99
Re-trained SR-HPM	83	68	0.51
Original SR-HPM	82	67	0.53

Table 6: *Correlation matrix of break types labeled by the re-trained SR-HPM (R) and the original SR-HPM (O).*

R\O	B0	B1	B2-1	B2-2	B2-3	B3	B4
B0	30287	1981	348	2	193	0	0
B1	1385	77972	798	154	1069	0	0
B2-1	414	971	19277	531	459	0	0
B2-2	0	181	514	**14217**	374	557	2
B2-3	186	1130	367	299	7801	0	0
B3	0	0	1	**1007**	1	**9987**	**849**
B4	0	0	0	1	0	**1017**	**8105**

Table 7: *RMSEs (ms) of the reconstructed pause duration.*

Break type	B0	B1	B2-1	B2-2	B2-3	B3	B4	total
Re-trained SR-HPM	2.4	18.5	24.4	83.1	29.1	100.6	143.9	50.6
Original SR-HPM	2.4	18.5	24.5	86.0	30.5	101.0	146.0	51.2

4.5. Evaluation of Prosody Generation

Table 8 shows the RMSEs of PAFs generated by the original SR-HPM and the re-trained SR-HPM. It is found that RMSEs of **sp**, *sd,* and *pd* are lower for the re-trained SR-HPM than ones for the original SR-HPM in the full prosody generation case (using predicted break). It is known that accuracy of break prediction can affect the prediction of PAFs. We, therefore, conduct a prosody generation experiment that the PAFs are predicted with the correct breaks (i.e., the correct prosodic structure). The result shows that the RMSEs of *sd* and **sp** of the re-trained SR-HPM are still lower than the ones of the original SR-HPM. We may partially conclude that the proposed hidden ISR estimation and the re-estimation of *sd* NF are effective in prosody modeling and prosody generation.

Table 8: *RMSEs of prosodic features generated by the original SR-HPM and the re-trained SRHPM with predicted and correct breaks.*

	SR-HPM		Re-trained SR-HPM	
	predicted break	correct break	predicted break	correct break
sd (ms)	49.1	48.2	48.8	47.7
sp (logHz)	0.1597	0.1472	0.1580	0.1467
se (dB)	3.63	3.54	3.63	3.53
pd (ms)	88.2	55.2	87.4	55.2

5. Conclusions

This paper proposes to estimate the so-called hidden inverse speaking rate (ISR) based on the existing SR-HPM. The hidden ISR is estimated by syllable durations being normalized by the factors of lexical information of tone and base syllable type, and prosodic structure. Several experiments were conducted to prove that the proposed hidden ISR is more meaningful and accurate to represent speaker's intended/underlying speaking rate than conventional raw speaking rate. In the future, we will apply the proposed ISR estimation approach to model variable or dynamic speaking rate for spontaneous speech.

6. Acknowledgements

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

7. 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.