A New Approach of Speaking Rate Modeling for Mandarin Speech Prosody

Chiao-Hua Hsieh\textsuperscript{1}, Chen-Yu Chiang\textsuperscript{1}, Yih-Ru Wang\textsuperscript{1}, Hsiu-Min Yu\textsuperscript{2}, Sin-Horng Chen\textsuperscript{1}

\textsuperscript{1}Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan
\textsuperscript{2}Language Center, Chung Hua University, Taiwan
kiwi0387@hotmail.com, {gene.cm91g@nctu, yrwang@mail.nctu, kuo@chu, schen@mail.nctu}.edu.tw

\section*{Abstract}
A new approach of Mandarin-speech prosody modeling to consider the effects of speaking rate is proposed. The approach is a modification of our previous prosody labeling and modeling method to take speaking rate as a continuous independent variable and let prosodic-acoustic features and some parameters of prosodic models depend on it in order to count its influences. A speaking rate-dependent hierarchical prosodic model is hence constructed from four speech corpora of a single female speaker with fast, normal, medium and slow speaking rates. An analysis of the effects of speaking rate on the model parameters showed that they agreed well with our prior knowledge. So, the proposed approach provides a systematic and effective way to quantify the effects of speaking rate on Mandarin-speech prosody.

\textbf{Index Terms}: speaking rate, prosody modeling

\section{1. Introduction}
Speaking rate is a prosodic feature that influences many speech phenomena such as syllable duration, pause duration, occurrence frequency of pause, and so on. Estimating the speaking rate of the speech signal [1] and exploring its effects on prosodic/linguistic features [2-4] are interesting research issues. Modeling the effects of speaking rate is also an important research issue in both automatic speech recognition (ASR) and text-to-speech (TTS). For ASR, the issue is to compensate the speaking rate effect for improving the low recognition performance of fast or slow speech [5-11]. For TTS, the speaking rate control of the synthetic speech is needed for making it sound more vividly to away from the criticism of machine-like sounding [12-14] as well as for being suitable for some special applications, e.g. fast rate for people with vision disability [15,16].

In this study, the effect of speaking rate on prosody modeling for Mandarin speech is explored. The proposed speaking rate modeling method is a modified version of the prosody labeling and modeling (PLM) method we proposed previously [3,18]. It is in line with our previous study of investigating the effect of speaking rate on Mandarin speech prosody via constructing individual hierarchical prosodic model (HPM) for four parallel corpora of a female speaker with fast, normal, medium and slow speaking rates [3]. Here, we take speaking rate as a continuous variable and construct a single HPM using the same four corpora. The study focuses on the effects of speaking rate on the prosodic-acoustic features and the parameters of the prosodic model. Fig. 1 shows the histogram (utterance count) of speaking rate (SR) in second/syllable of the four databases. All utterances are short paragraphs. There are in total 1,478 utterances consisting of 203,746 syllables. As shown in the figure that the SR of utterances in these four databases distribute in the range of 0.147-0.297 second/syllable (or 3.4-6.8 syllables/sec) and overlap seriously.

The paper is organized as follows. Section 2 gives a brief review of the HPM. Section 3 presents the proposed approach in detail. Experimental results are discussed in Section 4. Some conclusions and future works are given in the last section.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig1.png}
\caption{Histogram of utterance’s speaking rate of four databases used in the study.}
\end{figure}

\section{2. Review of the HPM}
The HPM [18] is a hierarchical prosodic model designed to describe the various relationships of prosodic-acoustic features, prosodic structure, and linguistic features. Three types of prosodic-acoustic features are considered. One is syllable-based features including syllable pitch contour represented by coefficients of a 3-rd order orthogonal polynomial expansion [18], \( sp_n = [a_0, a_1, a_2] \), syllable duration \( sd_n \), and syllable energy level \( se_n \) of the \( n \)-th syllable. Another is syllable-juncture features including pause duration \( pd_n \) and energy-dip level \( ed_n \) of juncture between the \( n \)-th and (\( n+1 \))-th syllables. The other is inter-syllable differential features including normalized pitch-level jump \( p_{jd} \), and two normalized duration lengthening factors, \( dl_n \) and \( df_n \), of syllable juncture \( n \). The complete prosodic-acoustic feature sequence is denoted as \( A = [X, Y, Z] \), where \( X = [sp, sd, se] \), \( Y = [pd, ed] \) and \( Z = [pj, dl, df] \).

The prosodic structure of the HPM is shown in Fig. 2. It is a modified version of the HPG model proposed by Tseng [19]. It is composed of four types of layered prosodic constituents: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath/prosodic phrase group (BG/PG). In the HPM, the prosody hierarchy is represented in terms of two types of prosody tags \( T = [B, P] \): the break type \( B \) of syllable juncture and the prosodic state \( P \) of syllable. The four prosodic constituents are delimited by seven break types denoted as \( B_0, B_1, B_2-1, B_2-2, B_2-3, B_3, \) and \( B_4 \) [18]. First, \( B_0 \) and \( B_1 \) represent non-breaks of reduced and normal boundary within a PW. Second, PW boundary \( B_2 = [B_2-1, B_2-2, B_2-3] \) is perceived as a minor-break boundary with F0 reset, short pause and duration lengthening, respectively. Third, PPh boundary \( B_3 \) has a clear pause. Fourth, \( B_4 \) is a breathing pause or a complete speech paragraph end. The prosodic state \( P \) is conceptually defined as the state in a prosodic phrase to account for the prosodic-acoustic feature variations imposed on higher-level prosodic constituents. Three types of prosodic states are used: \( p \) for pitch, \( q \) for duration and \( r \) for energy.
3. Speaking Rate Modeling

Fig. 3 shows a schematic diagram of the proposed speaking rate modeling approach for Mandarin speech prosody. For each utterance \( k \), the average syllable duration, \( \mu^{pd}_{k} \), with all pauses being excluded is first calculated and taken as a measure of its speaking rate \( SR(k) \). The prosodic-acoustic features of the utterance are then normalized by \( SR \)-specific normalization functions to compensate the effects of \( SR \). Lastly, a modified PLM algorithm, which incorporates the speaking-rate effects into the break-syntaic submodel, is conducted to simultaneously construct the HPM and label prosodic tags of the speech corpus.

3.1. Prosodic-Acoustic Feature Normalization

Conventionally, syllable durations are \( Z \)-normalized using utterance-based mean and standard deviation [17]. But, this may result in over-normalization, e.g., utterances of similar \( SR \) but with quite different standard deviations (see Fig. 4) resulted from some interesting factors other than speaking rate will be normalized to suppress those factors. In this paper, we adopt a more conservative approach to use a smooth curve to model the relation of the original normalization parameters (e.g. standard deviations of syllable duration) of utterances and their \( SR \)'s in the whole database. The \( SR \)-specific normalization functions for each utterance are then formed using the parameters calculated from the smoothed curve and used to compensate the effects of speaking rate on the prosodic-acoustic features.

3.1.1. Syllable duration normalization

Syllable durations of utterance \( k \) are Gaussian-normalized with mean \( \mu^{pd}_{k} = SR(k) \) and a smoothed standard deviation instead of the original one. Fig. 4 displays the scattering plots of syllable duration vs. \( SR \) (left) and the utterance-wise standard deviation of syllable duration vs. \( SR \) (right). It is found that the standard deviation increases as \( SR \) increases. By fitting the standard deviations of utterances with a second-order polynomial, the normalization function for syllable duration \( sd \) is expressed by

\[
sd' = \left( sd - \mu^{pd}_{k}\right) / \sigma^{pd}(SR(k)) \times \mu^{sd}_{k} + \sigma^{sd}_{k}
\]

where \( \mu^{sd}_{k} \) and \( \sigma^{sd}_{k} \) are global mean and standard deviation of syllable duration.

![Fig. 4: The scattering plots of \( sd \) vs. \( SR \) (left) and the utterance-wise standard deviation of \( sd \) vs. \( SR \) (right).](image)

3.1.2. Pause duration normalization

By observing the distribution of pause duration \( pd \), we find that Gamma distribution is suitable for its modeling, i.e.,

\[
G(pd; \alpha, \beta) = \beta^{x} \Gamma(\alpha)^{-1}(pd)^{\alpha-1}e^{-\beta pd}, \text{ for } x \geq 0 \text{ and } \alpha, \beta > 0
\]

Since the parameters of Gamma distribution, \( \alpha^{pd}_{k} \) and \( \beta^{pd}_{k} \), of utterance \( k \) can be represented in terms of its mean \( \mu^{pd}_{k} \) and standard deviation \( \sigma^{pd}_{k} \), we therefore calculate the smoothed mean \( \tilde{\mu}^{pd}(SR(k)) \) and standard deviation \( \tilde{\sigma}^{pd}(SR(k)) \) first, then calculate the smoothed \( \tilde{\alpha}^{pd}(SR(k)) \) and \( \tilde{\beta}^{pd}(SR(k)) \) from them, and last form SR-specific normalization function. Fig. 5 displays the scattering plots of \( \tilde{\mu}^{pd}_{k} \) vs. \( SR \) (left) and \( \tilde{\sigma}^{pd}_{k} \) vs. \( SR \) (right). It can be seen from the figure that both the mean and standard deviation of \( pd \) increase as \( SR \) increases. Two second-order polynomials are used to respectively model the relationships between \( \mu^{sd}_{k} \) and \( \sigma^{sd}_{k} \) vs. \( SR \), i.e.,

\[
\tilde{\mu}^{pd}(SR) = a_{1}(SR)^{2} + b_{1} \cdot SR + c_{1}
\]

(4)

\[
\tilde{\sigma}^{pd}(SR) = a_{2}(SR)^{2} + b_{2} \cdot SR + c_{2}
\]

(5)

Then, the observed \( pd \) is distribution-normalized by

\[
pd'' = G^{-1}(G(pd; \tilde{\alpha}^{pd}(SR(k)), \tilde{\beta}^{pd}(SR(k))) \cdot \alpha^{pd}_{k}, \beta^{pd}_{k})
\]

(6)

where \( G(pd; \alpha, \beta) \) is the CDF of Gamma distribution;

\[
\tilde{\alpha}^{pd}(SR(k)) = \left( \tilde{\mu}^{pd}(SR(k))^{2} \right)^{1/2}(\tilde{\sigma}^{pd}(SR(k))^{2})
\]

(7)

\[
\tilde{\beta}^{pd}(SR(k)) = \tilde{\mu}^{pd}(SR(k))^{2} / (\tilde{\sigma}^{pd}(SR(k))^{2})
\]

(8)

and \( \alpha^{pd}_{k} \) and \( \beta^{pd}_{k} \) are parameters calculated from global mean and standard deviation.
Fig. 6 shows the scattering plots of $pd$ and $pd'$ vs. $SR$ for three syllable-juncture types: intra-word, inter-word without punctuation mark (PM) (non-PM inter-word), and inter-word with PM (PM inter-word). As shown in these figures, the effects of speaking rate on $pd$ are properly normalized by the proposed method especially for the two inter-word cases.

Fig. 7: The scattering plots and fitting lines for the utterance-wise mean and standard deviation of $sp_{i}(2)$ of tone 4.

### 3.2. The modified PLM method

After feature normalization, a modified version of the PLM method proposed previous [17] is employed to automatically train a speaking rate-dependent HPM and label all utterances with the two types of prosodic tags: prosodic state and break type. In this study, we assume that the effects of speaking rate on the distributions of the prosodic-acoustic features have been generally removed by the normalization procedures illustrated in Section 3.1. Therefore all utterances of different speaking rates can be used in the modified PLM method to train the HPM and label the prosodic tags. But we still need to let some parameters of the HPM be speaking rate dependent in order to compensate its influences. Since the frequency of break is known to highly depend on the speaking rate [3], we consider it in this study by letting the break-syntax submodel be $SR$-dependent, i.e., $P(B_{m} | L_{n}, SR(k)) \equiv \{1 P(B_{m} | L_{n}, SR(k)) \}$.

This is realized by the following two steps. In Step 1, the marginal probability $P(B_{m} | L_{n})$ is estimated as in [18] from the labeled $B_{m}$ by the CART decision tree algorithm. In Step 2, the scattering plot of frequency vs. $SR$ for each break type in each leaf node of the decision tree constructed in Step 1 is formed and linearly fitted to obtain $P(B_{m} | L_{n}, SR)$, i.e.,

$$P(B_{m} = m | L_{n}, SR(k)) = \sum_{all break types} P(B_{m} = m | L_{n}, SR(k)) \approx \frac{c_{m,n} SR(k) + d_{m,j}}{\sum_{all break types} c_{j,n} SR(k) + d_{j,n}}$$

where $B_{m}$ is a break type of syllable juncture $n$ in utterance $k$; $j$ represents the leaf node index associated with the linguistic feature vector $L_{n}$; and $c_{m,n}$ and $d_{m,j}$ are linear regression coefficients for break type $m$ and leaf node $j$.

Fig. 8 displays two examples: one for short-pause minor break $B2-2$ in a node of non-PM inter-word and another for long-pause major break $B4$ in a PM node. It shows that the occurrence probability of $B2-2$ at a non-PM inter-word juncture is low for small $SR$ (fast speech) and high for large $SR$ (slow speech), while the dependency of the occurrence probability of $B4$ at a PM juncture on $SR$ keeps the same trend but with lower slope. These findings agree well with our prior knowledge about Mandarin speech prosody [3].

### 3.1.3. Syllable LogF0 contour normalization

LogF0 contour of each syllable is represented by four orthogonally-transformed parameters $sp_{i} = [a_{0i}, a_{1i}, a_{2i}, a_{3i}]^{T}$ [17,18]. By observing the scattering plots of the utterance-wise mean/std deviations of each coefficient vs. $SR$, we find that it is necessary to compensate the effects of speaking rate on syllable LogF0 contours for each lexical tone and for each dimension, i.e.,

$$sp_{i}(i) = sp_{i}(i) - \frac{\tilde{\mu}(SR(k),t_{s},i)}{\tilde{\sigma}(SR(k),t_{s},i)} \times \mu(t_{s},i) + \mu(t_{s},i)$$

where $\mu(R,S,R,i)$ and $\tilde{\sigma}(R,S,i)$ represent respectively the SR-specific normalization functions of mean and standard deviation of $sp$ for dimension $i$ and tone $t$; and $\mu(t_{s},i)$ and $\sigma(t_{s},i)$ are the global mean and standard deviation. In this study, two 1st order polynomials of $SR$ are adopted for $\mu(t_{s},i)$ and $\sigma(t_{s},i)$.

Fig. 7 shows an example of the scattering plots and fitting lines for the utterance-wise mean $\mu(t_{s},i)$ and standard deviation $\sigma(t_{s},i)$ of $sp_{i}(2)$ (i.e., slope $a_{0i}$) of tone 4. It can be seen from the figure that both mean and standard deviation are highly correlated with the speaking rate.

### 3.1.4. Syllable energy level normalization

By observing the scattering plots of syllable energy level, we find that they truly depend on the recording condition of each utterance but not speaking rate. Therefore, we simply let them be Gaussian-normalized to the global mean and standard deviation on an utterance-by-utterance basis.

### 4. Experimental Results

The modified PLM algorithm took 94 iterations to reach a convergence in total log-likelihood. Table 1 displays the total residual errors (TREs) of the three reconstructed prosodic-acoustic features using different combinations of affecting patterns (APs) with proper denormalization. Values in the table show the effects of removing the influences of some APs considered on $sp$, $sd$, and $se$. These results generally agreed with those achieved in our previous study [3].
Table 2: RMSEs (ms) of pause duration for different break types.

<table>
<thead>
<tr>
<th>Break type</th>
<th>RMSE (ms)</th>
<th>B0</th>
<th>B1</th>
<th>B2</th>
<th>B22</th>
<th>B23</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ tone</td>
<td>67.3</td>
<td>70.6</td>
<td>61.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ coarticulation</td>
<td>63.2</td>
<td>50.1</td>
<td>48.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ p</td>
<td>0.8</td>
<td>1.4</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pause duration is the most important feature in the syllable-juncture prosodic-acoustic submodel. Fig. 9 displays the means of pause duration vs. SR for 7 break types. They agreed well with our prior knowledge about break duration. Table 2 shows the RMSEs of the reconstructed pause duration for different break types. It can be found from the table that they were high only for B2-2, B3 and B4. Since these three break types were minor and major breaks, and were more tolerable for large modeling errors, these results were reasonable.

Fig. 9: Average pause durations of break types vs. SR. Values obtained in [3] by separately modeling four databases of different speaking rate are also shown in the figure.

Table 2: RMSEs (ms) of pause duration for different break types.

<table>
<thead>
<tr>
<th>Break type</th>
<th>RMSE (ms)</th>
<th>B0</th>
<th>B1</th>
<th>B2</th>
<th>B22</th>
<th>B23</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ tone</td>
<td>67.3</td>
<td>70.6</td>
<td>61.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ coarticulation</td>
<td>63.2</td>
<td>50.1</td>
<td>48.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ p</td>
<td>0.8</td>
<td>1.4</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 10 (left) shows the distributions of labeled breaks in non-PM inter-word syllable junctures. The most significant finding from the figure is that B2-2 (minor break of short pause) occurred more frequently in slower speaking rate. This agreed with the prior knowledge that speakers tend to insert more pauses within a sentence as they speaker slower. This also matches with the results shown in Fig. 10 (right) that the average length of prosodic word was shorter in an utterance of larger SR.

Fig. 10: (Left): The distributions of labeled breaks in non-PM inter-word syllable junctures. (Darker nodes represent higher probabilities) (Right): A scattering plot of average length of prosodic word in syllable vs. speaking rate. Values obtained in [3] are also shown in the figure.

5. Conclusions and Future Works

This paper proposed a new systematic approach for modeling the effects of speaking rate on Mandarin-speech prosody. Both the parameters of the HPM trained and the labeled prosodic tags were linguistically meaningful. In the near future, the proposed modeling approach will be applied to construct HPMs for assisting in ASR and speaking rate conversion in TTS.

6. Acknowledgements

This work was supported in part by the NSC of Taiwan under Contract No. NSC98-2221-E-009-075-MY3 and No. NSC99-2221-E-009-009-MY3; and in part by the MoE ATU plan.

7. References