Advanced Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech and Its Application to Prosody Generation for TTS

Chen-Yu Chiang, Sin-Horng Chen and Yih-Ru Wang

Dept. of Communication Engineering, National Chiao Tung University, Taiwan
gene.cm91g@nctu.edu.tw, schen@mail.nctu.edu.tw, yrwang@cc.nctu.edu.tw

Abstract
Motivated by the success of the unsupervised joint prosody labeling and modeling (UJPLM) method for Mandarin speech on modeling of syllable pitch contour in our previous study, in this paper, the advanced UJPLM (A-UJPLM) method is proposed based on UJPLM to jointly label prosodic tags and model syllable pitch contour, duration and energy level. Experimental results on the Sinica Treebank corpus showed that most prosodic tags labeled were linguistically meaningful and the model parameters estimated were interpretable and generally agreed with other previous study. In virtue of the functions given by the model parameters, an application of A-UJPLM to the prosody generation for Mandarin TTS is proposed. Experimental results showed that the proposed method performed well. Most predicted prosodic features matched well to their original counterparts. This also reconfirmed the effectiveness of the A-UJPLM method.

Index Terms: prosody modeling, prosody labeling, prosody generation, text-to-speech system

1. Introduction
Prosody generation plays a very important role on the naturalness of the synthesized speech in a TTS system. The main concern is to explore an appropriate mapping from the linguistic features extracted from the input text to the prosodic features of the synthesized speech. Concluding from the previous studies, prosody generation methods can be roughly divided into two classes, including (1) direct modeling approach and (2) hierarchical prosody approach. In direct modeling approach, for Mandarin speech, explicit numerical prosodic targets such as syllable or initial-final duration, syllable pitch contour, syllable energy level and inter-syllable pause duration are directly predicted from input linguistic features by some pattern recognition tools like artificial neural networks [1] and recurrent neural network [2], etc. The main advantage of this approach is that it can be automatically established without much help of linguistic expertise. However, it is very difficult to analyze and modify its hidden structure when some unsatisfactory prosody is generated. This approach usually focuses on modeling of prosodic feature variation in each syllable, initial or final rather than considering prosody hierarchy. On the other hand, in hierarchical prosody approach [3], symbolic prosodic features such as break types are first predicted from the input linguistic features. A prosody hierarchical structure, for example, from top layer to lower layer: intonation phrase, prosodic phrase, prosodic word, and syllable, is hence obtained from the predicted symbolic prosodic features. Then, numerical prosodic features are obtained by superimposing prosodic patterns of various layers pre-stored or generated from a model, or by selecting prosodic templates from a speech inventory. Through this approach, suprasegmental prosodic pattern and variation of each level can be easily analyzed and manipulated. However, it encountered some drawbacks of inconsistencies of break labeling caused by human's individual subjectivity and fatigue during long time labeling so as to propagate error to prosody modeling. It is also a time consuming task that requires rich linguistic expertise.

To circumvent the above-mentioned disadvantages, an unsupervised joint prosody labeling and modeling method (UJPLM) for Mandarin was proposed in the previous study [4], a new scheme intended to construct interpretable statistical prosodic models and to consistently label prosodic tags for an unlabeled Mandarin speech corpus. The task automatically determines prosodic tags that describe a prosody hierarchy and to build four prosodic models simultaneously. In this paper, motivated by the success of the UJPLM on the modeling of syllable pitch contour, we extend the study to include the other two important prosodic features, syllable duration and energy level. The study will jointly model syllable pitch contour, duration and energy level using the same method presented in the previous study. For simplicity, this extension is referred to as the advanced UJPLM (A-UJPLM) method. Then, an application of A-UJPLM to the prosody generation for Mandarin TTS is proposed. A more substantial prosody labeling and modeling for Mandarin speech was achieved by the A-UJPLM method. This paper is organized as follows. A brief review of UJPLM is given in Section 2. Section 3 introduces the proposed A-UJPLM in this work. Section 4 presents the proposed model-based approach to prosody generation. Some conclusions are drawn in the last section.

2. Review of UJPLM
2.1. Prosody Hierarchy and Prosodic Tags
Fig. 1 displays the prosody hierarchy adopted in the previous and current studies. It consists of four layers: syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breath/prosodic phrase group (BG/PG). Two types of prosody tags, break type and prosodic state of syllable, are employed to characterize the prosodic constituents of these four layers. A set of six break types, \( A = \{B0, B1, B2-1, B2-2, B3, B4\} \), is used to delimit these four prosody layers. Notice that B2-1 and B2-2 are divided from B2 and defined as a PW boundary with irregular F0 reset and perceived short pause, respectively.

![Fig. 1: A conceptual prosody hierarchy of Mandarin speech.](image)

The prosodic state tag is conceptually defined as the state in a prosodic phrase and used as a substitution for the effects from high-level linguistic features, such as a word, a phrase or a syntactic tree. It is also assumed to account for the prosodic variation contributed by higher-level prosodic constituents (i.e. PW, PPh and BG/PG) that carry suprasegmental prosodic
property. On the other hands, low-level affecting factors refer to some syllable-level linguistic features which represent intrinsic characteristics of Mandarin prosody on SYL level, such as lexical tones and base-syllable type.

2.2. Design of UJPLM

The UJPLM can be regarded as an unsupervised clustering problem that labels prosodic tag \((B^s, p^s)\) given prosodic features \((s_{p}, p_d, e_d)\) and linguistic features \((t_{n+1}^+, l)\):

\[
B^s_{n}, p^s_n = \arg\max_{B^s, p^s} \left[ \prod_{n=1}^{N} P(s_p_n | p_{d_n}, B^s_n, t^+_{n+1}) \prod_{n=1}^{N} P(p_d_n | p_{d_n}, B^s_n) \right] \times \prod_{n=1}^{N} P(p_d_n | p_{d_n}, B^s_n)
\]

(1)

where \(B^s_{n} = \{ B_n \}_{n=1}^{N} \) is the break tag sequence of an utterance with \(B_n \) being the break type of the inter-syllable location following \(n\)-th syllable (referred to juncture \(n\)); \(p^s_n = \{ p_n \}_{n=1}^{N} \) is the prosodic state tag sequence with \(p_n\) representing pitch prosodic state of \(n\)-th syllable; \(s_{p}\) is syllable pitch contour of \(n\)-th syllable represented by 3rd order orthogonal expansion coefficients; \(p_d\) and \(e_d\) represent respectively pause duration and energy-dip level of juncture \(n\); \(l\) denotes the contextual linguistic feature vector around juncture \(n\); \(t^+_{n+1}(t_{n+1}, t_{n+1})\) denotes tone sequence of previous, current and following tones; \(P(s_p_n | p_{d_n}, B^s_n, t^+_{n+1})\) is the syllable prosodic model to describe the variations in syllable pitch contours controlled by lexical tones and prosodic tags; \(P(p_{d_n | p_{d_n}, B^s_n})\) is the prosodic state model to describe the variation of prosodic state sequence controlled by the break type; \(P(p_{d_n, e_d} | B_n)\) represents the break-acoustics model to construct the relationship between the break type and acoustic features of pause duration and energy-dip level; \(P(B_n)\) is the break-syntax model to describes the relationship between the break type of a syllable juncture and contextual linguistic features.

The syllable pitch contour model is then considered to compute four additive affecting patterns (APs), i.e.

\[
P(s_p | p_{d}, B^s_{n+1}) = N(s_p | \beta_{n+1} + \beta_{n+1, n} + \beta_{n+1, n+1} + \mu, \sigma^2)
\]

(2)

where \(\beta_{n+1}\), \(\beta_{n+1, n}\), \(\beta_{n+1, n+1}\) and \(\beta_{n+1, n+1}\) are APs of tone, prosodic state, forward and backward tone coarticulations respectively. Via introducing this additive model, the variation in pitch contour is elegantly decomposed into the effect of higher-level prosodic constituents by \(\beta_{n+1}\), and the effect of SYL level by \(\beta_{n+1, n}\). Notice that the coarticulation effect is elegantly treated to consider different degrees of coupling between two neighboring syllables via letting it depend on the break type of the syllable juncture. The break-acoustics model is expressed by the product of a Gamma distribution for pause duration and a normal distribution for energy dip. Both the break-acoustics and the break-syntax models are elaborated by decision trees with question sets formed by contextual linguistic feature vector \(l\). Then, with proper initialization of the break types and prosodic state, the UJPLM model is trained by a special designed sequential optimization procedure. After well training, all parameters of the four models are obtained and the whole database is properly labeled with the two types of prosody tags, i.e., break type and prosodic state of syllable pitch level.

3. Advanced UJPLM

In the A-UJPLM method, a new prosodic model to jointly consider the modeling of syllable pitch contour, duration, and energy level is proposed. We discuss the new prosodic model in detail in the following subsections.

3.1. Features and Parameters Used in A-UJPLM

Aside from the prosodic features used in the UJPLM of pitch modeling, we consider more features in this study, including syllable duration \(sd_n\) syllable energy level \(se_n\). As for prosodic tags, two new types of prosodic states, the duration prosodic state \(q_n\) and the energy prosodic state \(r_n\), are introduced for the modeling of syllable duration and energy level to consider the effects contributed from high-level prosodic constituents of PW, PPh and PG/BG. Besides, a new break type \(B_2\)-3 is added to represent the syllabic boundary of \(B_2\) with perceived lengthening of the preceding syllable. Concerning modeling of syllable duration and energy level, we add effecting factors of base-syllable \(s_n\) and final types \(f_n\) because they are two important syllable-level linguistic features, other than syllable tone, that seriously affect the variations of syllable duration and energy level. An utterance-level normalization factor \(u_n\) is added to consider respectively the variation in syllable duration due to the speaking rate and the variation in syllable energy level due to the recording volume.

3.2. Design of A-UJPLM

The A-UJPLM is a modified version of UJPLM extended with syllable duration model \(P(sd_q, r_n, s_n, u_n)\), syllable energy model \(P(se_q, r_n, t_n, f_n, u_n)\), duration prosodic state model \(P(q|p_{d}, B_n)\), and energy prosodic state model \(P(r|p_{d}, B_n)\):

\[
P(B^s_n, p^{s}_n, q^{n}, r^{n}, u^{n}) = \arg\max_{B^s_n, p^{s}_n, q^{n}, r^{n}, u^{n}} \left[ \prod_{n=1}^{N} P(s_p_n | p_{d_n}, B^s_n, t^+_{n+1}) \prod_{n=1}^{N} P(p_d_n | p_{d_n}, B^s_n) \right] \times \prod_{n=1}^{N} P(p_d_n | p_{d_n}, B^s_n)
\]

(3)

Syllable duration and energy models can be elaborated by

\[
p(sd_q | q_n, t_n, f_n, u_n) = N(sd | \gamma_{q_n} + \gamma_{t_n} + \gamma_{f_n} + \mu, \sigma^2)
\]

(4)

\[
p(se_q | r_n, t_n, f_n, u_n) = N(se | \alpha_{r_n} + \alpha_{t_n} + \alpha_{f_n} + \mu, \sigma^2)
\]

(5)

where \(\gamma\)’s and \(\alpha\)’s are APs for syllable duration and energy models; \(\mu\) and \(\sigma\) denote respectively the global mean and the variance of residual. The A-UJPLM is also trained by the sequential optimization procedure adopted in UJPLM. After well training, all model’s parameters are obtained and the database is properly labeled with break type, and pitch/duration/energy prosodic states.

3.3. Experimental Results on A-UJPLM

An unlabeled read Mandarin speech database containing 425 utterances with 56237 syllables uttered by a female professional announcer was used to evaluate the proposed A-UJPLM. The database is divided into two parts: a training set of 379 utterances with 52192 syllables for construction of the prosodic models and a test set of 46 utterances with 4801 syllables for evaluation of the proposed prosody generation method. In this experiment, the numbers of pitch, duration, and energy prosodic states were all empirically set to be 16.

Table 1 displays the total residual errors (TREs, defined as the percentage of sum-squared residue over the observed sum-squared) of the modelings for \(s_{p}, sd_n\), and \(se_n\) with respect to
the use of different combinations of APs. It can be seen from the table that the TREs reduced as more APs were considered and the most significant one is prosodic state. The variations of syllable prosodic features of $sp_n$, $sd_n$, and $se_n$ were effectively removed by the considered APs.

### Table 1: TREs (% of the modelings for $sp_n$, $sd_n$, and $se_n$ w.r.t. the use of different combinations of APs.

<table>
<thead>
<tr>
<th>Pitch</th>
<th>Duration</th>
<th>Energy level</th>
</tr>
</thead>
<tbody>
<tr>
<td>APs</td>
<td>TRE</td>
<td>APs</td>
</tr>
<tr>
<td></td>
<td>+ Utterance</td>
<td>+ Utterance</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>+ Tone</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>+ Coarticulation</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>+ Prosodic state</td>
</tr>
</tbody>
</table>

The means of pause duration for seven break types of B0, B1, B2-1, B2-2, B2-3, B3 and B4 were 3, 11, 18, 109, 16, 287 and 543 ms, respectively. This result confirmed that the break types of higher level were generally associated with longer pause duration. The structure of the constructed decision tree for the break-syntax models generally agreed with the one constructed by UJPLM in the previous study [4]. The APs of pitch, duration and energy prosodic states spanned widely to cover the whole dynamic ranges of syllable log-F0 level, duration and energy level variations with lower indices of prosodic state representing lower log-F0 levels, shorter syllable durations, and lower energy level respectively. In the constructed prosodic state models, low to high pitch and energy prosodic state index jumps were clearly observed on the juncture of B3 and B4 illustrating pitch and energy resets in the beginning of a PPh or BG/PG while high to low duration prosodic state index jumps on the juncture of B3 and B4 illustrating apparent pre-boundary lengthening effect in the end of a PPh or BG/PG.

A typical example of the labeling results by A-UJPLM is given in Fig. 2. Most breaks labeled were the same except for an inserted B2-3 after the 5th syllable and a substitution of B3 with B4 after the 11th syllable. The insertion of B2-3 seemed to be reasonable because there existed an apparent syllable duration lengthening on the 5th syllable. For each prosodic feature of syllable log-F0 mean, duration and energy level, the curve formed by integrating the prosodic-state APs and global mean showed smoother PW patterns as compared with those of the observed zigzag curve. The last syllables of all PWs had longer syllable duration illustrating the pre-boundary duration lengthening effect. It is also found that apparent resets existed on the energy prosodic state of the last syllables of most PWs manifesting clear stress patterns. These phenomena matched well with the findings of Tseng [5].

The proposed A-UJPLM was able to construct interpretive prosodic models and to generate proper prosodic tags automatically. Therefore, it can be directly used or extended to be used in the application of prosody generation for TTS system as presented in Section 4.

### 4. A-UJPLM-based Prosody Generation

In this section, an A-UJPLM-based approach is proposed for prosody generation. It is composed of two steps: break prediction and prosodic feature prediction. In the break prediction step, a break type sequence is predicted for each input text by the break-syntax model $P(B_t | l_s)$ using some linguistic features extracted from the input text and plays an important role to properly parse the input text into strings of PWs, PPhs, and BG/PG. In the prosodic feature prediction step, four types of prosodic features, including syllable pitch contour, syllable duration, syllable energy level, and intersyllable pause duration, are generated from input linguistic features and the break-type sequence generated in the first step by using the syllable pitch/duration/energy/ prosodic models, the prosodic state models, and the break-acoustic model.

To evaluate the performance of the proposed break prediction and prosodic feature prediction methods, the test set of the Sinica Treebank corpus is adopted. It is labeled in advance by the prosodic models trained in Section 3 with seven types of break and three types of prosodic states by decoding of Eq. (3).

#### 4.1. Break Prediction

Based on the decision tree of break-syntax model trained in section 3, we let the decision tree grow using more sophisticated linguistic features to predict break only from linguistic feature. A window up to six words is adopted in this study to extract the POS and word length features for the break prediction of the current word juncture: three words before and after the juncture. Table 2 displays the prediction accuracy. It can be found from the table that the prediction rates were high for B1 and B4, medium for B3 and B0, and low for B2-1, B2-2 and B2-3. By analysis of the confusion matrix between predicted and correct breaks, it is found that minor breaks (B2-1/2-2/2-3) were easily confused with non-break and major break were easily predicted as B2-2. For improving the prediction rate, especially minor break, utilizing more sophisticated linguistic features, such as syntactic phrase or syntactic tree is worthwhile doing in the future.

### Table 2: Prediction accuracy (% of (a) three broad classes of break, and (b) the seven break types.

<table>
<thead>
<tr>
<th></th>
<th>(a) Non-break</th>
<th>Minor break</th>
<th>Major break</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92.4</td>
<td>77.9</td>
<td>82.0</td>
<td>88.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b)</th>
<th>B0</th>
<th>B1</th>
<th>B2-1</th>
<th>B2-2</th>
<th>B2-3</th>
<th>B3</th>
<th>B4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>70.6</td>
<td>86.3</td>
<td>56.4</td>
<td>52.3</td>
<td>31.8</td>
<td>61.0</td>
<td>89.5</td>
</tr>
</tbody>
</table>

#### 4.2. Prosodic Feature Prediction

The prosodic features to be predicted include syllable prosodic features ($sp_n$, $sd_n$, $se_n$) and inter-syllable pause duration ($pd_n$). Among them, the inter-syllable pause duration of each syllable juncture can be simply predicted by the break-acoustic model trained in section 3, i.e.

$$pd_n = \arg \max_{pd} P(pd | B_n, l_s)$$

where $B_n$ represents the optimal break type of juncture $n$ predicted by the break-syntax model predicted in Section 4.1. The syllable prosodic features, including syllable pitch contour $sp_n$, syllable duration $sd_n$ and syllable energy level $se_n$ are predicted by the models formulated basing on the minimum mean squared error (MMSE) criterion. Given with
the predicted break sequence \( B' \) and linguistic feature sequence \( l \), the MMSE predictors for \( sp, sd, \) and \( se \) are:

\[
\hat{sp}_{1} = E[sp(B_1'|B_1)], \quad \hat{sd}_{1} = E[sd(B_1'|B_1)], \quad \text{and} \quad \hat{se}_{1} = E[se(B_1'|B_1)],
\]

respectively. Since \( sp, sd, \) and \( se \) are predicted in the same way, we only present the prediction model of \( sp \) here for simplicity. The MMSE predictor for \( sp \) can be elaborated by

\[
\hat{sp}_{1} = E[sp(B_1'|B_1)] = \int sp \cdot P(sp|B_1') \, dp \quad \text{by Eq. (7)}
\]

\[
= \sum_{p} \left( \beta'_{1} + \beta_{2} + \beta'_{3} + \beta_{4} + \beta'_{5} \right) P(p|B_1') \, dp, \quad (7)
\]

It can be seen from Eq. (7) that the predicted syllable pitch contour is a weighted sum of the reconstructed patterns formed by superimposing various APs with weights being the \( a \ posteriori \) probabilities of prosodic state \( p \). The \( a \ posteriori \) probability \( P(p|B_1') \) can be calculated by

\[
P(p|B_1') = \frac{P(p, B_1')}{\sum_{p} P(p, B_1')} \quad n=1
\]

\[
\quad \sum_{p} P(p_{n-1}, B_{n-1}, l_{n}) P(p_{n}, B_{n}|B_1') \quad 2 \leq n \leq N \quad (8)
\]

where \( P(p_{n-1}, B_{n-1}, l_{n}) \), which is a modified version of the prosodic state model \( P(p_{n-1}, B_{n-1}) \), strengthens the influences of linguistic features and the breaks \( B_1 \) on the current prosodic state \( p \). Notice that Eq. (8) is very similar to the calculation of the forward probability in hidden Markov model with the initialization (\( n = 1 \)) and induction (\( 2 \leq n \leq N \)).

In practical realization, since the space of the histories \( \{ p_{n-1}, B_{n-1}, l_{n} \} \) and linguistic features \( \{ l \} \) is too large, we partition the space into several classes \( C(p_{n-1}, B_{n-1}, l_{n}) \) to calculate the conditional probabilities \( P(p_{n} | C(p_{n-1}, B_{n-1}, l_{n})) \) by the decision tree method.

Table 3(a) displays the TREs of the prosody prediction results for syllable pitch contour, duration and energy level. The performances were acceptable. To separate the effect of break prediction on the prosodic feature prediction, we do the same experiment using the correct break labels. Table 3(b) displays the experimental results. By comparing the results shown in the two tables, we find that the latter performed better. This shows that the break prediction plays an important role in the prediction of prosodic features. Erroneous breaks predicted will make gross shifts of PW patterns and may result in large prediction errors of prosodic features. Hence, to improve the prosodic feature generation, the break prediction task is essential and worthwhile further investigating in the future.

Table 3: TREs of the prosodic feature prediction results using (a) predicted break (b) correct break labels

<table>
<thead>
<tr>
<th></th>
<th>( sp )</th>
<th>( sd )</th>
<th>( se )</th>
<th>( pd )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>42.7</td>
<td>46.2</td>
<td>36.0</td>
<td>18.9</td>
</tr>
<tr>
<td>(b)</td>
<td>39.1</td>
<td>41.7</td>
<td>33.3</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Fig. 3 displays an example of the predicted prosodic features by the A-UJPLM-based approach. It illustrates the prosodic feature variations of two sentential utterances extracted from a long utterance. It can be found from the figure that the predicted prosodic features matched well with their original counterparts for most syllables. Some large errors can be found to occur on the syllable durations and inter-syllable pause durations of the first sentence. They were mainly resulted from a series of break prediction errors. For example, the two contiguous break prediction errors (predict \( B_2-2, B_1 \) as \( B_1, B_3 \)) in the first sentence caused a gross shift of the first PW showing a move of the phrase-ending lengthening from the 4th syllable to the 6th syllable. The after-phrase long pause also shifted two syllables to the right synchronously. By using correct break labels, these large prosodic feature prediction errors disappeared. This confirmed that break prediction errors are responsible for prosodic feature prediction errors.

5. Conclusions

In this paper, an advanced unsupervised joint prosody labeling and modeling method for Mandarin speech is proposed to jointly model syllable pitch contour, duration and energy level. Experimental results showed that the prosodic models constructed and the prosodic tags labeled were linguistically or prosodically interpretable and meaningful. The model parameters can be directly used or elaborated to be utilized in prosody generation for TTS system. So it is a very promising prosody modeling method for Mandarin speech.

6. Acknowledgements

This work was supported by NSC under contract NSC 96-2218-E-002-001. The authors also want to thank Academia Sinica, Taiwan for providing the Treebank Corpus.

7. References