An Automatic Prosody Labeling Method for Mandarin Speech

Chen-Yu Chiang¹, Hsiu-Min Yu², Yih-Ru Wang¹ and Sin-Horng Chen¹

¹Dept. of Communication Engineering, National Chiao Tung University, Taiwan
²Dept. of Foreign Languages and Literature, Chung Hua University, Taiwan
gene.cm91g@nctu.edu.tw, Kuo@chu.edu.tw, yrwang@cc.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract
A new model-based automatic prosody labeling method for Mandarin speech is proposed. It first introduces four models to describe the relationships of the prosody tags to be labeled, the prosodic features of the speech signals, and the linguistic features of the associated texts. It then employs a sequential optimization procedure to estimate parameters of these four models and find all prosody tags. Experimental results on the Sinica Tree-Bank corpus showed that most prosody tags labeled were meaningful and the estimated parameters of these four models matched well with our a priori knowledge about Mandarin prosody.

Index Terms: Prosody labeling, Mandarin prosody

1. Introduction
Automatic prosody labeling is to detect important prosody cues (e.g. breaks) from input speech signal and the associated text. A conventional pattern recognition approach [1,2] is to first use a well-annotated speech corpus to learn the relationship between prosody tags to be labeled and some prosodic features extracted from the speech signal in the training phase; and to then classify prosody tags from the prosodic features of the testing signal. Besides, linguistic features, such as word boundary, POS and syntactic phrase boundary, are used to set constraints to assist in the prosody classification. The main drawback of the approach is the lack of consistency in human’s labeling because a lot of acoustic and linguistic factors should be considered. Moreover, it is a labor-intensive task when the corpus is large.

To circumvent these drawbacks, a new model-based automatic prosody labeling method for Mandarin speech is proposed in this paper. It first introduces four models to describe the relationships of the prosody tags to be labeled, the prosodic features of the speech signals, and the linguistic features of the associated texts. It then employs a sequential optimization procedure to estimate parameters of these four models and find all prosody tags. Two advantages of the method can be found: (1) No human-labeled training data is needed, and (2) Prosody labeling and modeling are accomplished simultaneously.

The paper is organized as follows. Section 2 briefly describes the prosody structure of Mandarin speech. Section 3 presents the proposed method. Experimental results are discussed in Section 4. Some conclusions are drawn in the last section.

2. Prosody Structure of Mandarin Speech
Fig. 1 displays a conceptual prosody hierarchy of Mandarin speech. In the structure, syllables (SYL) located in the lowest layer are the basic prosody units to form prosodic words (PWs) of the second layer. PWs are then concatenated to form minor prosodic phrases (MIPPHs) of the third layer. MIPPHs are combined to form major prosodic phrases (MPPHs) in the top layer.

\[
B = \{B_0, B_1, B_2, B_3, B_4\}
\]

Fig. 1: A conceptual prosody hierarchy of Mandarin speech.

Following the guideline of ToBI labeling system [3], five break types, B0–B4, of inter-syllable pause are defined. Their relationships with the prosody hierarchy are shown in Fig.1. B0 represents an intra-PW syllabic boundary that the two neighboring syllables are tightly coupled. B1 is also an intra-PW syllabic boundary but with normal coupling. B2 represents an inter-PW syllabic boundary with short pause or minor pitch reset. B2 is further divided into two subclasses: B2-1 (with minor pitch reset) and B2-2 (with apparent short pause). B3 and B4 represent minor and major breaks with medium and long pauses, respectively. Besides, they usually accompany medium or large pitch resets.

3. The Proposed Method
The task of the prosody labeling is to determine the break types of inter-syllable pauses and the prosodic states of syllables given with prosodic features, including syllable pitch contour features, pause duration, and pause energy dip, as well as linguistic features of various levels. Here, prosodic state is a tag to conceptually represent the state of the syllable in a prosodic phrase. The proposed model-based approach is to perform the prosody labeling via assuming four models describing the relationship of those features. The task can be formulated as a parametric optimization problem:

\[
B^*, P^* = \arg\max_{B, P} P(B, P, S, P, D, P, E, L, T)
\]

where \(B = \{B_0, B_1, B_2, B_3, B_4\}\) and \(B_{n,a} \in [B0–B4]\) represents the break type of the inter-syllable pause following syllable \(n\) in utterance \(k\) (referred to as pause \((k,n)\) thereafter); \(N_s\) is the total number of syllables in utterance \(k\); \(K\) is the total number of utterances; \(P = \{p_{n,a} | n = 1 \cdots N_s; k = 1 \cdots K\}\) and \(p_{n,a} \in [1–P]\) is the prosodic state of syllable \(n\) in utterance \(k\) (syllable \((k,n)\) thereafter), and \(P\) is the total number of prosodic states; \(SP = \{sp_{n,a}\}\) and \(sp_{n,a}\) is the vector of four orthogonal expansion coefficients representing the observed pitch contours of syllable \((k,n)\); \(PD = \{pd_{n,a}\}\) and \(pd_{n,a}\) is the duration of pause \((k,n)\); \(PE = \{pe_{n,a}\}\) and \(pe_{n,a}\) is the energy dip of pause \((k,n)\); \(L = \{l_{n,a}\}\) and \(l_{n,a}\) is the vector of contextual linguistic features around pause \((k,n)\); \(T = \{t_{n,a}\}\) and \(t_{n,a}\) is the tone of


494 August 27–31, Antwerp, Belgium
syllable \(k,p\). \(P(\text{SP,PD,PE,B,P,L,T})\) is the prosody model describing the variations of prosodic features \((\text{SP,PD,PE})\) controlled by the prosodic phrase structure elements \((B,P)\) and the syntactic phrase structure features \((L,T)\), and \(P(\text{B,P,L,T})\) is a prosody-syntax model which describes the relation of \((B,P)\) and \((L,T)\).

To make the problem mathematically tractable, \(P(\text{SP,PD,PE,B,P,L,T})\) is simplified by

\[
P(\text{SP,PD,PE,B,P,L,T}) \approx P(\text{B,P,T})P(\text{PD,PE,B,L})
\]

\[
\approx \prod_{x=1}^{N} \left[ P(\text{sp}_{x} | \text{pa}_{x},B_{x-1},B_{x},a_{x-1},a_{x},k_{x-1},k_{x},l_{x}) \right] P(\text{pd}_{x} | \text{pa}_{x},a_{x})
\]

where \(P(\text{sp}_{x} | \text{pa}_{x},B_{x-1},B_{x},a_{x-1},a_{x},k_{x-1},k_{x},l_{x})\) is a syllable pitch contour model which describes the dependence of \(\text{sp}_{x}\) on the nearby prosody structure cues (prosody tags) and tones; \(P(\text{pd}_{x} | \text{pa}_{x},a_{x})\) is a pause acoustic model which describes the dependence of \(\text{pd}_{x}\) and \(\text{pe}_{x}\) on \(B_{x}\) and some nearby linguistic features \(l_{x}\).

Similarly, we simplify \(P(\text{B,P,L,T})\) by

\[
P(\text{B,P,L,T}) = P(\text{B,L})P(\text{P,L})P(\text{B,L}) = P(\text{P,L})P(\text{B,L})
\]

\[
\approx \prod_{x=1}^{N} \left[ P(\text{pa}_{x}) \prod_{x=1}^{N} P(\text{pa}_{x},B_{x},a_{x}) \prod_{x=1}^{N} P(B_{x},l_{x}) \right]
\]

where \(P(\text{pa}_{x})\) is the initial prosodic state probability; \(P(\text{pa}_{x},B_{x-1},B_{x})\) is the prosodic state transition probability, and \(P(B_{x},l_{x})\) is a break-syntax model which describes the dependence of \(B_{x}\) on some nearby linguistic features \(l_{x}\). In this study, the break-syntax model is trained by the decision tree method.

\[
P(\text{sp}_{x} | k_{x},p_{x},B_{x-1},B_{x},a_{x-1},a_{x},k_{x-1},k_{x},l_{x})\]

is then elaborated to

\[
\text{sp}_{x} = \text{sp}_{0} + \beta_{k_{x},p_{x}} + \beta_{a_{x-1},a_{x}} + \beta_{a_{x},p_{x}} + \mu
\]

where \(\text{sp}_{0}\) is the normalized (i.e., residual) pitch contour; \(\beta_{k_{x},p_{x}}\) and \(\beta_{a_{x},p_{x}}\) are the affecting patterns of the tone and prosody state, respectively; \(\beta_{a_{x-1},a_{x}}\) and \(\beta_{a_{x},p_{x}}\) are the forward and backward coarticulation affecting patterns; \(\mu\) is the mean of the global mean pattern. It is noted that \(\beta_{k_{x},p_{x}}\) is set to have nonzero value only in its first dimension for restricting the influence from prosodic state merely on the pitch level of the current syllable. Fig. 2 displays the relationship of \(\text{sp}_{x}\) and these affecting factors.

![Fig. 2: The relationship of sp_x and its affecting factors.](image)

By assuming that \(\text{sp}_{x}\) is normally distributed, i.e.,

\[
N(\text{sp}_{x};\mu,\sigma^{2})
\]

we have

\[
P(\text{sp}_{x} | \text{pa}_{x},B_{x-1},B_{x},a_{x-1},a_{x},k_{x-1},k_{x},l_{x}) = N(\text{sp}_{x};\mu_{\text{pa}_{x}},\sigma_{\text{pa}_{x}}^{2})
\]

and

\[
P(\text{pd}_{x} | \text{pa}_{x},a_{x})
\]

is also simplified and expressed by the product of a Gamma distribution for pause duration and a normal distribution for energy dip:

\[
P(\text{pd}_{x} | \text{pa}_{x},a_{x}) = \frac{1}{\Gamma(\alpha_{\text{pd}_{x}})} \left( \beta_{\text{pd}_{x}} \right)^{\alpha_{\text{pd}_{x}}} \exp\left( -\beta_{\text{pd}_{x}} \right)
\]

In this study, the pause acoustic model is separately trained by the decision tree method for each break type.

3.1. Linguistic Features

Linguistic features around the break are extracted for constructing decision trees of pause acoustic model and break-syntax model. The features come from various levels including syllable (e.g. initial types, inter/intra-word), word (e.g. punctuation mark, word length, specific words, POS), syntactic tree level (e.g. size/boundary of syntactic phrase) and sentence level (e.g. length of a sentence). There are totally 57 and 222 linguistic features for pause acoustic and break-syntax models respectively.

3.2. The Training of the Models

To perform prosody labeling and estimate the parameters of these four models, a sequential optimization procedure based on the ML criterion is adopted. It first defines a likelihood function by

\[
Q = \log \prod_{x=1}^{N} \left[ \prod_{x=1}^{N} P(\text{sp}_{x} | \text{pa}_{x},B_{x-1},B_{x},a_{x-1},a_{x},k_{x-1},k_{x},l_{x}) \prod_{x=1}^{N} P(\text{pd}_{x} | \text{pa}_{x},a_{x}) \prod_{x=1}^{N} P(B_{x},l_{x}) \prod_{x=1}^{N} P(\text{pa}_{x}) \right]
\]

Then, with proper initialization, it sequentially performs the following operations to optimize \(Q\): (1) Update the affecting patterns of tone, coarticulation, and prosodic state; (2) Relabel the prosodic states and break types; and (3) Construct the decision trees to update the parameters of the pause acoustic model and the break-syntax model. Specifically, the training procedure executes the following steps until a convergence is reached:

**Step 0: Initialization**

- Calculate \(\mu\) by averaging pitch contours of all syllables.
- Calculate \(\{\beta_{i}; t=1\sim5\}\) by averaging all residue pitch contours \(\text{sp}_{x}\) of each tone.
- Label break types heuristically by the following decision tree using pause duration and energy.

![Fig. 3: The decision tree for initial break type labeling](image)
4. Experimental Results

The proposed method was evaluated using a Mandarin speech database. The database contained the read speech of a female professional announcer. Its texts were all short paragraphs composed of several sentences selected from the Sinica Tree-Bank Corpus [4]. The database consisted of 380 utterances with 52192 syllables. All syllable segmentation and F0 detection were first done automatically using HTK and ESPS, respectively, and then error corrected manually. In the prosody labeling, we set the number of prosodic states to 16. The optimization process converged after 30 iterations.

In the following, we analyze the experimental results.

4.1. The Syllable Pitch Contour Model P(SP,B,P,T)

After well training, the covariance matrices of the original and normalized syllable logF0 were

\[
\begin{bmatrix}
87.9 & 24.0 & -23.6 & -4.5 \\
24.0 & 90.5 & 9.6 & -4.2 \\
-23.6 & 9.6 & 17.8 & -4.8 \\
-4.5 & -4.2 & -4.8 & 5.0
\end{bmatrix} \quad \text{and} \quad
\begin{bmatrix}
9.6 & 0.5 & -0.5 & 0.0 \\
0.5 & 31.9 & 2.7 & -1.5 \\
-0.5 & 2.7 & 11.1 & 6.6 \\
0.0 & -1.5 & 0.6 & 3.7
\end{bmatrix}
\]

This showed that the influences of many affecting factors have been removed. Fig.4 displays the patterns of five tones (left) and prosodic states (right). The learned tone patterns matched very well with our knowledge of standard tone patterns.

Fig.4: Tone patterns (left) and prosodic state patterns (right).

Fig.5 shows coarticulation patterns \(p_{t,r}^{B0/B1} / p_{t,r}^{B3/B4}\) for B0/B1/B4. It can be observed that many forward (backward) patterns were bent in their beginning (ending) parts. This mainly resulted from the pitch level mismatch between the beginning and ending of the preceding-current (current-following) F0 contours; e.g. \(p_{t,r}^{B0/B1} | p = (1,2),(1,3),(1,5),(2,2),(2,3)\) were due to H-L mismatches, while \(p_{t,r}^{B3/B4} | p = (3,1),(3,4),(5,1)\) corresponded to L-H mismatches. We also observed that the curvatures of patterns for B0 were larger than those for B1 and B4. This confirms that the lower level of break type corresponds to the case of more tightly coupling between the two consecutive syllable F0 contours. The most interesting patterns are \(p_{t,r}^{B0/B1}\) for B0 and B1 which go upward drastically to conform to the well-known 3-3 tone sandhi rule. Lastly, most forward (backward) patterns for B4 were upward (downward). This exhibits the reset (offset) phenomenon in the beginning (end) syllable of a MPPH.

Fig.5: Forward (up)/backward (down) coarticulation patterns \(p_{t,r}^{B0/B1}\) for B0(point line)/B1(solid line)/B4(dash line). Here \(tp = (i,j)\)

4.2. Pause Acoustic Model P(PD,P,E,L)

Fig. 6 displays the distributions of pause duration and energy dip for different break types. It can be found from the figure that break type of higher level were associated with longer pause duration and lower energy dip.

Fig. 6: The pdf of pause duration (left) and energy dip (right). Numbers in ( ) denote the mean values.

Fig. 7 shows the decision trees of pause acoustic model for B4/B3/B1/B0. Generally, the pause acoustic model of higher-level break type is affected by higher-level linguistic features. For B4 and B3, PM and sentence/phrase size are affecting factors. For B1 and B0, initial types of fricative and stop are affecting factors.

Fig. 7: Decision trees of \(P(pd_{t,r},pe_{t,r},B_{t,r},l_{t,r})\) for B4/B3/B1/B0. The numbers in the brackets denote average pause duration in ms (left) and energy dip in dB (right) of the nodes.
4.3. Prosodic Transition Probabilities \( P(B_{kn}|B_{kp}, B_{lm}) \)

Fig. 8 displays the diagram of significant prosodic state transitions for different break types. For B0 and B1, high-to-low state transitions showed that the syllable pitch level declined within prosodic words. For B2, B3 and B4, low-to-high state transitions showed that the syllable pitch level reset across prosodic words and phrases. Obviously, higher-level break type corresponded to larger pitch-level reset.

Fig. 8: The diagram of significant prosodic state transitions.

4.4. Break-syntax Model \( P(B_{kn}|B_{kn}) \)

Fig. 9 shows the decision tree of the break-syntax model. The tree was divided into four sub-trees, T3-T6, by the three questions of \{PM\}, \{Minor PM\} and \{Type 1 intra-word\}. Here, Type-1 intra-word means within a short lexical word. Can be seen from the bar plots that T4, which corresponded to major PM, was mainly composed of B4. Similarly, T3 (i.e., minor PM) was mainly composed of B3 and B4, and T5 (Type 1 intra-word) was mainly composed of B0 and B1. Via further tracing T5, we find that it was likely labeled as B0 if the next syllable began with a null initial or initial in \{m, n, l, r\}, otherwise it was labeled as B1. These observations match well with our priori knowledge about inter-syllable pause.

Fig. 9: The decision tree of break-syntax model. The bar plot below a node denotes the distribution of break types (B0-B4) and the number is the total samples number of the node.

Fig. 10 displays the details of T6. A juncture whose next syllable has null initial or initial in \{m, n, l, r\} was mainly labeled as B0. A juncture associated with syntactic phrase boundary was more likely labeled as B2 or B3. If the following word is a “DE”, it was mainly labeled as B1 or B0.

Fig. 10: The details of Sub-tree T6.

4.5. A Labeling Example

Fig. 11 displays an example of the automatic prosody labeling. (a) syntactic trees with prosody tags and (b) syllable logF0 means: observed (open circle) and prosodic state+global mean(close circle). Solid/dash/dot lines represent B3/B2-1/B2-2 respectively.

Fig. 11: An example of the automatic prosodic labeling. (a) syntactic trees with prosody tags and (b) syllable logF0 means: observed (open circle) and prosodic state+global mean(close circle). Solid/dash/dot lines represent B3/B2-1/B2-2 respectively.

5. Conclusions

A new model-based method of automatic prosody labeling for Mandarin speech has been discussed. Its efficiency has been confirmed by experiment using the Sinica Tree Bank Corpus. Applications to ASR and TTS as well as the extension of the method to incorporate the syllable duration and energy level models are worth further studying.

6. Acknowledgements

This work was supported in part by NSC under contract NSC95-2218-E-002-027 and NSC95-2752-E-009-014-PAE. The authors also want to thank Academia Sinica, Taiwan for providing the Tree-Bank Corpus.

7. References