LATENT PROSODY MODEL OF CONTINUOUS MANDARIN SPEECH

Chen-Yu Chiang[1], *Xiao-Dong Wang*[2], *Yuan-Fu Liao*[3], *Yih-Ru Wang*[1],
Sin-Horng Chen[1], *Keikichi Hirose*[4]

[1]Dept. of Communication Engineering, National Chiao Tung University, Taiwan
[2]Department of Electronic Engineering, University of Tokyo, Japan
[3]Department of Electronic Engineering, National Taipei University of Technology, Taiwan
[4]Department of Information and Communication Engineering, University of Tokyo, Japan

ABSTRACT

The major difficulty of prosody modeling and automatic tone recognition of continuous Mandarin speech is the complex interaction of tones and prosody/intonation on F0 contours. In this study, we propose a latent prosody model (LPM) aiming to jointly model the affections of tone and prosody state on F0. The main purposes are twofold including (1) automatic prosody state labeling and (2) improving tone recognition accuracy. The basic idea is to introduce latent prosody state variables into an additive statistic model of F0 which already considers the affecting factors of tone and speaker. Experiments on the Tree-Bank corpus showed that LPM not only gave meaningful prosody state labeling results but also improved the average tone recognition rate from 80.86% of a multi-layer perceptron (MLP) baseline to 82.55%.

Index Terms**—**speech processing, speech recognition, tone recognition

1. INTRODUCTION

As a tonal language, in Mandarin speech, there is a tight interaction between four lexical tones (Tones 1~4) and a neutral tone (Tone 5), and the underlying speech prosody/intonation. Many works on prosody and tone modeling have been reported [1-4], where prosody and tone are usually modeled separately. To treat them separately, it is necessary to remove effects from tones when modeling prosody, and, in turn, it is necessary to suppress prosody when recognizing syllable tone types.

Taking the conventional tone recognition approach as an example, tone recognition methods could be roughly divided into three groups including (1) feature normalization, (2) contextual features/context-dependent models and (3) prosody model front-end. One good example of the feature normalization is the tone nucleus model [1-2] which defines the sentential F0s as the concatenation of target point, tone nuclei and transition loci. The contextual feature/context-dependent models [3] want to capture differential features between neighboring syllables (i.e., inter-syllable features) for assisting tone recognition. This kind of approaches usually comes along with the context-dependent tone models. The prosody model front-end-based scheme aims on detecting the prosody states in the front-ends to assist a backend tone recognizer. Multilayer perceptron (MLP) and recurrent neural network (RNN)-based prosodic modeling [4] have been reported.

However, these three classes of tone recognition schemes all separately model the affections of tone and prosody on F0 and use two-stage approaches. In this paper, we propose a latent prosody model (LPM) aiming to jointly model the affections of tone and prosody state on F0. The main purposes are twofold including (1) automatic prosody state labeling and (2) improving tone recognition accuracy at the same time.

The idea is to introduce additional latent prosody state variables into an additive statistical model of F0 which already considers the affecting factors of tone and speaker. In LPM, we use syllable as the basic modeling units, and the distribution of the observed F0 contour of a syllable in an utterance is formulated as follows:

$$\mathbf{x}_{k,n} = \mathbf{y}_{k,n} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\gamma}_{p_{k,n}} + \boldsymbol{\mu} \qquad (1)$$

where $\mathbf{x}_{k,n}$ and $\mathbf{y}_{k,n}$ are two-dimensional vectors (mean and slope of log$F$0 contour of a syllable) representing, respectively, the observed and normalized pitch contours of the n-th syllable in utterance k; $\boldsymbol{\mu}$, $\boldsymbol{\chi}_{t_{k,n}}$ and $\boldsymbol{\gamma}_{p_{k,n}}$ are the affecting factor of speaker, tone $t_{k,n} \in \{1,2,3,4,5\}$ and prosody state $p_{k,n} \in \{1,2,\cdots,P\}$, respectively.

The normalized pitch contour $\mathbf{y}_{k,n}$ is further modeled using a Gaussian distribution $N(\mathbf{y}_{k,n};\mathbf{0},\mathbf{R})$, or equivalently the observed pitch contour $\mathbf{x}_{k,n}$ is modeled by

$$P(\mathbf{x}_{k,n} \mid t_{k,n}, p_{k,n}, \lambda) = N(\mathbf{x}_{k,n}; \boldsymbol{\mu} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\gamma}_{p_{k,n}}, \mathbf{R}) \qquad (2)$$

By this joint modeling, the performance of both the automatic prosody state labeling and tone recognition of input utterances may then be improved. The LPM will be optimized using the maximum likelihood (ML) criterion and solved by the expectation-maximization (EM) training algorithm. Moreover, a well-trained LPM could be used online to simultaneously label the prosody states and tones of input utterances.

The paper is organized as follows. Section 2 presents an advanced LPM which additionally considering break and position-dependent state transition probabilities and describes the maximum likelihood-based LPM training algorithm. Section 3 proposes an online version of the LPM for both automatic prosody state labeling and tone recognition. Section 4 shows experimental results. Some conclusions are given in the last section.

2. ADVANCED LPM AND TRAINING ALGORITHM

2.1. Break/Position-Dependent Transition Probabilities

Speech is usually organized as prosodic phrase groups and the pitch contour usually declines within a prosodic phrase.

Considering this phenomenon, break type (long pause between two prosodic phrases) and state position (position in a prosodic phrase)-dependent state transition probabilities are also modeled and incorporated into the LPM. They are expressed as follows:

$$P(p_{k,n} \mid p_{k,n-1}, B_{k,n-1}) \tag{3}$$

where $B_{k,n-1} \in (MB, MIB, BP, MP, EP)$, MB and MIB are, respectively, the major and minor breaks between two prosody states $(p_{k,n-1}, p_{k,n})$, BP (beginning of a prosodic phrase), MP (middle of a prosodic phrase) and EP (end of a prosodic phrase) are the positions of a state in a prosodic phrase.

The types of the breaks are simply determined by the length of the pause duration between two prosodic phrases/words. Two thresholds are set, i.e., the pause duration of the *MIB*s and *MB*s should be larger than 100 ms and 300 ms, respectively.

All breaks and positions could be automatically detected from the segmentation information of an input utterance.

2.2. LPM Training Algorithms

To estimate the parameters of the LPM, a sequential optimization procedure based on the ML criterion is adopted. It first defines a likelihood function expressed by

$$L=\log\left\{\prod_{k=1}^{K}\left[\prod_{n=1}^{N_k}P(\mathbf{x}_{k,n}|t_{k,n},p_{k,n},\lambda)P(p_{k,1})\prod_{n=2}^{N_k}P(p_{k,n}|p_{k,n-1},B_{k,n-1})\right]\right\} \tag{4}$$

where K is the total number of utterances; and N_k is the total number of syllables in utterance *k*.

Then, the training procedure sequentially updates the two types of affecting factors (i.e., tone, prosody state), and re-labels the prosody state to optimize the likelihood function L. The sequential optimization training procedure executes the following steps until a convergence has reached. It is worth noting that each step updates a subset of LPM parameters.

Step 0: Initialization

- Calculate the speaker affecting factor $\boldsymbol{\mu}$ by averaging the feature vectors of all log $F0$ contour.
- Derive the initial affecting factors $\boldsymbol{\chi}_t$ of five tones by averaging all residue pitch $\mathbf{x}1_{k,n} = \mathbf{x}_{k,n} - \boldsymbol{\mu}$ of each tone.
- Derive the initial prosody state factors $\boldsymbol{\gamma}_p$ and label the prosody state of each syllable by vector quantization (VQ) using the residue pitch $\mathbf{x}2_{k,n} = \mathbf{x}_{k,n} - \boldsymbol{\chi}_{t_{k,n}} - \boldsymbol{\mu}$.
- Derive the initial covariance matrix $\mathbf{R}$
- Derive the initial prosody state transition probabilities using the statistics of labeled prosody states.

Step 1: Update the affecting factors $\boldsymbol{\chi}_t$ of five tones with all other parameters fixed.

Step 2: Re-label the prosody state sequence of all utterance using a Viterbi search algorithm to maximize L, i,e,,

$$\mathbf{p}_k^* = \underset{\mathbf{p}_k}{\arg\max}\log\left\{\left[\prod_{n=1}^{N_k}P(\mathbf{x}_{k,n} \mid t_{k,n}, p_{k,n}, \lambda)\right]\left[P(p_{k,1})\prod_{n=2}^{N_k}P(p_{k,n} \mid p_{k,n-1}, B_{k,n-1})\right]\right\} \tag{5}$$

And then update the affecting factors $\boldsymbol{\gamma}_p$ of P prosody states, the covariance matrix $\mathbf{R}$, and the prosody state transition probabilities.

Step 3: Repeat step 1 to 2 until convergence

3. ONLINE LPM

A well trained LPM could be exercised online to label both tone and prosody state sequences of an input test utterance. The proposed online LPM is shown in Fig. 1. It has three components including (1) a MLP tone recognizer (2) a LPM-based prosody state labeler and (3) a LPM-assisted MLP tone recognizer. It is worth noting that the second MLP utilizes the prosody state cues to improve tone recognition performance.

The operation of the online LPM is as follows. First, the MLP tone recognizer provides the initial guess of the tone sequence of an input test utterance. Then an iterative procedure is operated, i.e., the LPM-based prosody state labeler and LPM-assisted MLP tone recognizer update the labeling of tones and prosody states in turn, until a convergence is reached. During the iteration, both the tone recognizer and prosody state labeler output *a posteriori* probabilities during the iteration.

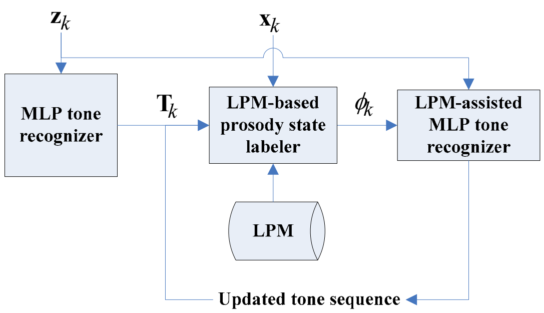

Fig. 1: The proposed block diagram of the online LPM for simultaneous prosody state detection and tone recognition.

3.1. LPM-based Prosody State Labeler

The output *a posteriori* probabilities of prosody state $\phi_{k,n}(i) = P(p_{k,n} = i \mid \mathbf{x}_k, \lambda)$ are derived using the following equations. It is similar to the forward/backward algorithm used to train hidden Markov models (HMMs):

$$P(p_{k,n}=i \mid \mathbf{x}_k,\lambda)=\frac{P(p_{k,n}=i,\mathbf{x}_k,\lambda)}{\sum_{j}P(p_{k,n}=j,\mathbf{x}_k,\lambda)}=\frac{\alpha_{k,n}(i)\beta_{k,n}(i)}{\sum_{j}\alpha_{k,n}(j)\beta_{k,n}(j)} \tag{6}$$

$$\begin{cases}\alpha_{k,n}(i)=P(\mathbf{x}_{k,1}\cdots\mathbf{x}_{k,n},p_{k,n}=i \mid \lambda)\\ \beta_{k,n}(i)=P(\mathbf{x}_{k,n+1}\cdots\mathbf{x}_{k,N} \mid p_{k,n}=i,\lambda)\end{cases}$$

3.2. LPM-Assisted MLP Tone Recognizer

The input feature vector of the LPM-assisted MLP tone recognizer includes (1) 16 *a posteriori* probabilities of prosody states $\phi_{k,n}(i), i=1\sim P$ given by LPM-based prosody state labeler and (2) 14 tone features $\mathbf{z}_{k,n}$ extracted from the underlying syllable. The 14 tone features are:

- log$F$0 mean, log$F$0 slope and energy mean of three uniformly segmented voiced part of the current syllable.
- pause duration between the current and the preceding/following syllables
- syllable duration
- a flag which represents the current syllable is at the beginning of an utterance or not
- a flag which represents the current syllable is at the end of an utterance or not

The LPM-based MLP tone recognizers have 3 layers. The output layer consists of five nodes, each corresponding to one of the five tones. Both the hidden and output layers use a standard sigmoid function. The training algorithm for the MLP is the back-propagation algorithm.

4. EXPERIMENTAL RESULTS

Performance of the LPMs, especially, the online one was evaluated using a Mandarin speech database. The database contained the read speech of a single female. Its texts were all paragraphs composed of several long sentences selected from the Sinica Tree-Bank corpus [5]. The database consisted of 380 utterances with 52192 syllables. The first 38 sentences (4775 syllables) were taken as test set; the rest 342 sentences (47,417 syllables) were taken as training set.

4.1. Constructed LPM

In the training phase, we set the numbers of prosody states to be 16. After the LPM was well trained, the covariance matrices of the original and normalized syllable $F0$ for the training set are measured and shown as follows:

$$\mathbf{R}_{\mathbf{x}} = \begin{bmatrix} 611.39 & 0 \\ 0 & 7.71 \end{bmatrix} \times 10^{-4} \Rightarrow \quad \mathbf{R}_{\mathbf{y}} = \begin{bmatrix} 14.59 & 0 \\ 0 & 3.94 \end{bmatrix} \times 10^{-4}$$

Also, for the testing set, they are

$$\mathbf{R}_{\mathbf{x}} = \begin{bmatrix} 614.12 & 0 \\ 0 & 7.67 \end{bmatrix} \times 10^{-4} \Rightarrow \quad \mathbf{R}_{\mathbf{y}} = \begin{bmatrix} 15.17 & 0 \\ 0 & 3.92 \end{bmatrix} \times 10^{-4}$$

It is found that the variances in both training and testing sets had been significantly reduced by LPM. This is especially true for the pitch mean. Table 1 displays the learned affecting factors of five tones. They match well to the canonical pitch contours of the five Mandarin tones.

Table 2 shows the learned affecting factors of the 16 prosody states. The 16 states are sorted by the values of the affecting factors of prosody states (from low to high). Fig. 2 displays the distribution of prosody states at different prosodic phrase positions. It is found that the state groups (12, 13, 14, 15) and (4, 5, 7) mainly are located in the BP and the EP respectively. This distribution matches the characteristic of a prosodic phrase on F0 that a prosodic phrase begins with high pitch level and ends with low pitch values.

To investigate the relationship between prosody states and the structure of prosodic phrases in more detail, we further analyze the prosody state transition probabilities $P(p_{k,n} \mid p_{k,n-1}, B_{k,n-1})$ and the major transition paths of prosody states are shown in Fig. 3.

From Fig. 3, it could be found that the state transitions 13-14 and 12-13 which represent the rising pitch patterns are very frequent in the BP. The state transitions 14-13, 13-12 and 12-11 in the BP represent falling pitch patters. These two patterns together form a rising-falling pattern in the BP. If we further trace the state transition paths, for example, starting from state 12, the major path would be 12-13-12-11-9-10-7-5-4-3 and the pitch contour of this path is similar to a rising-falling slow variant F0 contour of a prosodic phrase.

It is also observed that prosody state 3 and 4 were usually located at the end of a prosodic phrase and state transitions 3-13 and 4-12 represent a F0 reset phenomenon accompanied a major break. If a prosodic phrase ends with prosody states 7 or 9, the state transitions 7-13 and 9-12 which represent $F0$ reset are frequent when there is a minor break.

Table 1: The learned affecting factors of the five tones.

Tone	1	2	3	4	5
Mean (LogF0)	0.17	-0.09	-0.21	0.09	-0.13
Slope	0.00	0.01	-0.03	-0.03	-0.01

Table 2: The learned affecting factors of the 16 prosody states.

state	1	2	3	4	5	6	7	8
LogF0	-0.85	-0.58	-0.41	-0.29	-0.2	-0.14	-0.09	-0.04
state	9	10	11	12	13	14	15	16
LogF0	-0.01	0.04	0.08	0.14	0.21	0.29	0.40	0.55

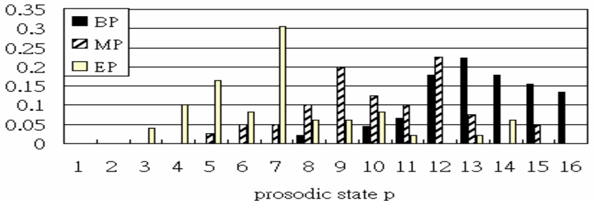

Fig. 2: The distribution of prosody states at different phrase positions.

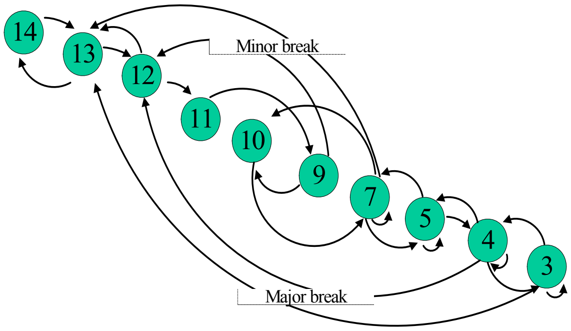

Fig. 3: Diagram of major prosody state transition paths inferred from the well trained LPM.

4.2. Online LPM Experiments

The capacities of the well trained LPM were evaluated on the online automatic prosody states labeling and tone recognition tasks

4.2.1. Automatic Prosody State Labeling

Fig. 4 displays a typical example of prosody state labeling in the testing set labeled by equation 5. The triangles represent original logF0 means of syllables normalized by the affecting factor of speaker mean and the circles represent the pitch means of the prosody state sequence labeled by the LPM. The dashed lines represent prosodic phrase boundaries.

It was observed that the circle lines between prosodic boundaries were smoother than the triangle lines. So the logF0 values of the prosody state sequence matched the nature of prosodic phrases, since the pitch contour of a prosodic phrase is a slow variant logF0 component. From the example, the prosody state jumps are also observed while there are prosodic boundaries (major/minor breaks).

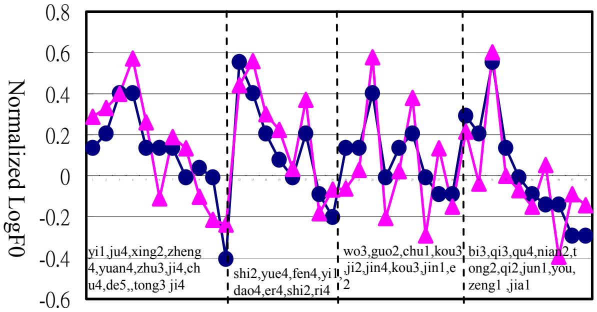

Fig. 4: A typical result of the LPM-based automatic labeling of prosody states of an input utterance (triangles: original logF0 means of syllables normalized by the speaker affecting factor, circles: pitch means of the corresponding prosody state sequence, dashed lines: correct prosodic phrase boundaries).

4.2.2. Automatic Tone Recognition

The number of nodes in the hidden layer of both the MLP and LPM-assisted MLP tone recognizers were empirically set to be 50. The confusion matrixes of the two tone recognizers were shown in Table 3 and 4.

The results show that the recognition rates of tone 1, 2, 3, 5 all have been improved, although the recognition rate of tone 4 degrade a little bit. The average recognition rate was improved from 80.86% to 82.55%. Especially for tone 3 and tone 5, the LPM-assisted MLP brought absolute 5.80% and 6.80% improvements comparing with the conventional MLP tone recognizer baseline, respectively.

From those tables, it is also shown that the tone 3 and tone 4 are often confused in the conventional MLP tone recognizer. The reason is that they may have the same log*F*0 slope and mean at different prosodic phrase positions. However, this problem could be partially solved by the LPM-assisted MLP, since the *a posteriori* probabilities of prosody states provide the extra information about the underlying prosodic status. The situation is similar for the recognition of tone 5 and 4 using the conventional tone recognizer.

Table 3: Confusion matrix of the conventional MLP tone recognizer

Ans\Rec	tone 1	tone 2	tone 3	tone 4	tone 5	Total
tone 1	88.17%	5.22%	0.75%	5.12%	0.75%	
tone 2	8.24%	84.89%	2.66%	2.75%	1.46%	
tone 3	5.19%	8.90%	56.00%	26.82%	3.09%	
tone 4	3.35%	1.92%	3.41%	90.20%	1.12%	
tone 5	4.80%	12.00%	7.20%	21.20%	54.80%	80.86%

Table 4: Confusion matrix of the LPM-assisted MLP tone recognizer

Ans\Rec	tone 1	tone 2	tone 3	tone 4	tone 5	Total
tone 1	90.19%	5.01%	0.75%	3.52%	0.53%	
tone 2	6.61%	86.78%	2.40%	1.80%	2.40%	
tone 3	3.96%	9.52%	61.80%	19.90%	4.82%	
tone 4	4.22%	1.49%	4.03%	88.72%	1.55%	
tone 5	5.60%	12.00%	7.20%	13.60%	61.60%	82.55%

5. CONCLUSIONS

A latent prosody model of Mandarin speech was developed in this paper. Experimental results on Tree-Bank corpus showed that the LPM not only could automatically label the prosody state but also improve tone recognition accuracy. Comparing with a MLP-based tone recognition baseline, the average recognition rate was improved from 80.86% to 82.55%. It is also found that the prosody state information provided by LPM is especially useful for classifying tone 3 and 5 from other types.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] J.-S. Zhang and K. Hirose, “Tone Nucleus Modeling for Chinese Lexical Tone Recognition”, *Speech Communication*, vol. 42, no. 4, pp. 447-466, 2004.

[2] J.-S. Zhang, S. Nakamura, and K. Hirose, “Tone Nucleus-based Multi-level Robust Acoustic Tonal Modeling of Sentential F0 Variations for Chinese Continuous Speech Tone Recognition”, *Speech Communication*, vol. 46, no. 4, pp. 440-454, 2005.

[3] W.-Y. Lin and L.-S. Lee, “Improved Tone Recognition for Fluent Mandarin Speech Based on New Inter-Syllabic Features and Robust Pitch Extraction,” *IEEE 8th Automatic Speech Recognition and Understanding Workshop*, PP.237-242, 2003

[4] Y.-R. Wang and S.-H. Chen, "Tone Recognition of Continuous Mandarin Speech Assisted with the Prosodic Model," J. Acoust. Soc. Am., Vol. 96(5), Pt.1, pp. 2637-2645, Nov. 1994.

[5] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao and K.-Y. Chen, “Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface”, *Proceedings of 2nd Chinese Language Processing Workshop* 2000, Hong Kong, pp. 29-37.