Latent Prosody Model-Assisted Mandarin Accent Identification

Yuan-Fu Liao[1], Shuan-Chen Yeh[2], Ming-Feng Tsai[3],

Wei-Hsiung Ting[4], and Sen-Chia Chang[5]

[1,2,3,4]Department of Electronic Engineering, National Taipei University of Technology
[5]Advanced Technology Center, Information and Communications Research Laboratories,
Industrial Technology Research Institute
[1,2,3,4]yfliao@ntut.edu.tw, [5]chang@itri.org.tw

Abstract

A two-stage latent prosody model-language model (LPM-LM)-based approach is proposed to identify two Mandarin accent types spoken by native speakers in Mainland China and Taiwan. The frontend LPM tokenizes and jointly models the affections of speaker, tone and prosody state of an utterance. The backend LM takes the decoded prosody state sequences and builds n-grams to model the prosodic differences of the two accent types. Experimental results on a mixed TRSC and MAT database showed that fusion of the proposed LPM-LM with a SDC/GMM+PPR-LM+UPR-LM baseline system could further reduced the average accent identification error rate from 20.7% to 16.2%. Therefore, the proposed LPM-LM method is a promising approach.

Keywords: Accent recognition, latent prosody model, Mandarin, Taiwan

1. Introduction

Over the past decades, many approaches have been proposed to deal with language identification (LID) tasks. They tried to capture the specific characteristics of different languages. These characteristics roughly fall into three categories: the phonetic repertoire, the phonotactics, and the prosody. The mainstream system (as shown in NIST language recognition evaluation (LRE) 2007) [1] is usually based on the fusion of multiple acoustic and phonotactic systems.

Although LID is extensively studied, less works have been done on accent identification (AID), especially for native speakers, such as American and Indian English, Mainland China and Taiwan Mandarin, Hindi and Urdu Hindustani and Caribbean and non-Caribbean Spanish. Comparing with LID task, AID of native speakers is more challenging because, (1) some linguistic knowledge, such as syllable structure, may be of little use since native speakers seldom make such mistakes; (2) difference among those speakers is relatively smaller than that among foreign (non-native) speakers. In other words, the capacities of the popular acoustic and phonotactic approaches may be limited in this case.

Many approaches have been proposed to model the prosodic differences between languages, dialects or accents [2], recently. Most of them are based on direct modeling of surface prosodic features, i.e., the raw prosodic features. For example, frame-level pitch flux features and GMMs were proposed in [3]; segmental-level pitch features were extracted using Legendre polynomials and modeled by ergodic Markov model in [4]; and supra-segment-level prosodic features were captured by n-gram in [5].

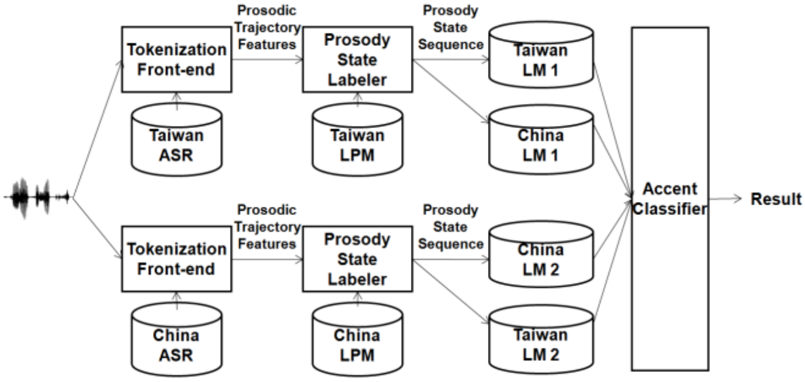

Figure 1. The block diagram of the proposed LPM-LM-based Mandarin accent identification system.

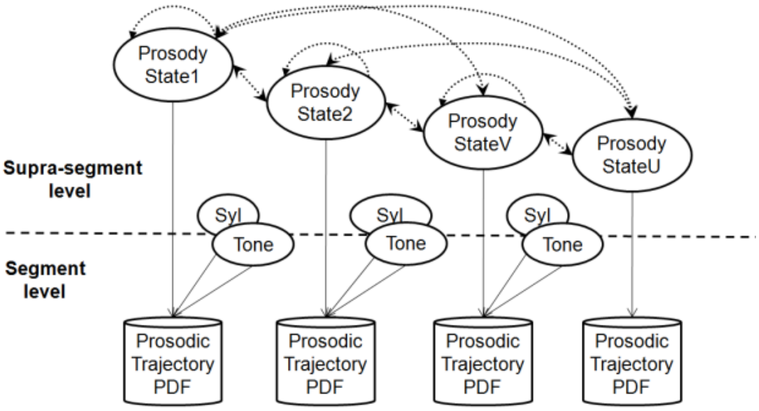

Figure 2. The block diagram of the proposed LPM framework (speaker factor is omitted to simply this figure).

However, surface prosodic features are often affected by many other non-prosodic latent factors, such as channel, speaker, phonetic context, and so on. Therefore, it is necessary to apply some feature normalization methods [6] to alleviate the unwanted affections. To absorb those unwanted affections, in this study a two-stage latent prosody model-language model (LPM-LM)-based approach as shown in Fig. 1 and 2 is proposed. The aim is to discriminate two Mandarin accent types spoken by native speakers in Mainland China and Taiwan.

In this approach, the frontend LPM [7] tokenizes (with the help of automatic speech recognizers (ASRs)) an input utterance into smaller prosodic units (sub-syllable in our case) and artificially introduces latent prosody states to represent the prosodic status of each token in an utterance. It then jointly models the affections of speaker, tone and prosody state on surface prosodic features in order to decode more precise prosody state sequences of the utterance. The backend LM then takes the decoded prosody state sequences and builds an n-gram to model the supra-segmental prosodic charactistics of each accent type.

In more detail, LPM as shown in Fig. 2 (1) introduces a two-level hierarchical structure of speech prosody [8] with prosodic states and state transition probabilities and (2) describes the joint affections of latent factors in a state by a variable-parameter probability density function whose parameters varies as a function of those latent factor-dependent parameters. The purpose is to explain the variant due to speaker, phonetic context and, especially, tone factors.

It is worth noting that (1) the proposed LPM-LM framework is similar to the popular parallel phone recognizer (PPR)-LM approach. However, the phone recognizers are replaced by automatic prosodic state tokenizers/labelers and, especially, (2) the LPM module could be trained in an unsupervised way to avoid any human annotation efforts.

This paper is organized as follows. Section 2 reviews the LPM framework. Section 3 discusses the application of LPM-LM on Mandarin AID. Section 4 reports the experimental results on a Mainland China and Taiwan Mandarin corpus. Some conclusions are given in the last section.

2. Latent Prosody Model of Speech Prosody

Based on the proposed LPM framework shown in Fig. 2, an input training utterance is first tokenized into a sequence of smaller prosodic units (sub-syllable in this case) including voiced and unvoiced segments. For each token, a segment-level prosodic feature vector $\mathbf{x}_n$ is extracted (coefficients of log-pitch and log-energy trajectories and the duration of the segment). Here, the coefficients of trajectories are computed using Legendre polynomial function from the raw log-pitch and log-energy contours. The speech prosody of an input utterance is thus represented by a sequence of segment-level prosodic feature vectors, i.e., $\mathbf{X}=\{\mathbf{x}_n, n=1,...,N\}$.

To well explain the variant of the observed prosodic feature vector sequence $\mathbf{X}$ of the utterance, several latent factors are introduced including speaker s, tone $\mathbf{T}=\{t_n, n=1,...,N\}$ (or major/minor stress in toneless language) and prosody state sequence $\mathbf{Q}=\{q_n, n=1,...,N\}$ (phonetic context is ignored in this study). The probability of $\mathbf{X}$ is defined as follows:

$$p(\mathbf{X})=\sum_{s,\mathbf{Q},\mathbf{T}} p(\mathbf{X}|s,\mathbf{T},\mathbf{Q})\,p(s,\mathbf{T},\mathbf{Q}) \tag{1}$$

Assume that each observed $\mathbf{x}_n$ is dependent only on local prosodic state q_n and tone t_n (and the speaker s), the first term in the right hand side of Eq. (1) is approximated as follows:

$$p(\mathbf{X}|s,\mathbf{T},\mathbf{Q})=\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n) \tag{2}$$

Assume that speaker, prosodic state and tone sequences are all independent variables and the probabilities of speaker s and tone sequence $\mathbf{T}$ are uniform distributions, the last term in the right hand side of Eq. (1) is approximated as follows:

$$p(s,\mathbf{T},\mathbf{Q}) \propto p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1}) \tag{3}$$

Finally, the distribution of the surface prosodic feature vector $\mathbf{x}_n$ is modeled by the following linearly additive [9] formulation:

$$\mathbf{x}_n = \mathbf{y}_n + \boldsymbol{\mu}_s + \boldsymbol{\mu}_{t_n} + \boldsymbol{\mu}_{q_n} \tag{4}$$

where $\mathbf{y}_n$ are prosodic feature vectors representing the normalized prosodic contours of the n-th syllable in an utterance; $\boldsymbol{\mu}_s$, $\boldsymbol{\mu}_{t_n}$ and $\boldsymbol{\mu}_{q_n}$ are the contributions of speaker s, prosody state q_n and tone t_n, respectively. The normalized pitch contour $\mathbf{y}_n$ is approximated using a zero mean Gaussian distribution $N(\mathbf{y}_n;\mathbf{0},\boldsymbol{\Sigma})$ (where $\boldsymbol{\Sigma}$ is diagonal matrix), or equivalently the observed prosodic feature vector $\mathbf{x}_n$ is modeled by

$$p(\mathbf{x}_n|s,t_n,q_n)=\mathbb{N}\left(\mathbf{x}_n;\boldsymbol{\mu}_s+\boldsymbol{\mu}_{t_n}+\boldsymbol{\mu}_{q_n},\boldsymbol{\Sigma}\right) \tag{5}$$

By this way, the likelihood function of an utterance given an LPM λ is expressed by

$$L(\mathbf{X}|\lambda)=\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n)\cdot p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1}) \tag{6}$$

Moreover, the optimal prosody state sequence $\hat{\mathbf{Q}}$ of an utterance could be automatically labeled using a Viterbi search algorithm (with or without tone tags given) which maximize the likelihood function $L(\mathbf{X}|\lambda)$, i.e.,

$$\hat{\mathbf{Q}}=\arg\max_{\mathbf{Q}} \log\left\{\prod_{n=1}^{N} p(\mathbf{x}_n|s,t_n,q_n)\cdot p(q_1)\prod_{n=2}^{N} p(q_n|q_{n-1})\right\} \tag{7}$$

3. LPM-based Mandarin Accent Identification

Mandarin spoken in Taiwan exhibits several major prosody differences from the Mandarin spoken in Mainland China [10]. Especially, people from Taiwan usually speak slower with a lower voice, and they sound soft and gentle; while Mainlanders have more ups and downs in their intonation, and their voices are higher and faster. These characteristics are likely attributable, at least in part, to influence from the Southern Fujianese dialect widely spoken throughout Taiwan.

Since there are prosodic differences between Mainlander’s and Taiwanese Mandarin, a LPM-based accent identification approach is built to identify these two Mandarin accent types. In the following subsections, the tokenization front-end and the speaker normalization parts of the proposed LPM-based approach and its training procedure are described in detail.

3.1. Tokenization front-end

The operation of the tokenization front-end is shown in Fig. 3. It firstly extracts the raw prosodic contours (log-pitch and log-energy) of an input utterance. The pitch and energy contours are then segmented by an ASR engine. The output is a sequence of voiced and unvoiced segments.

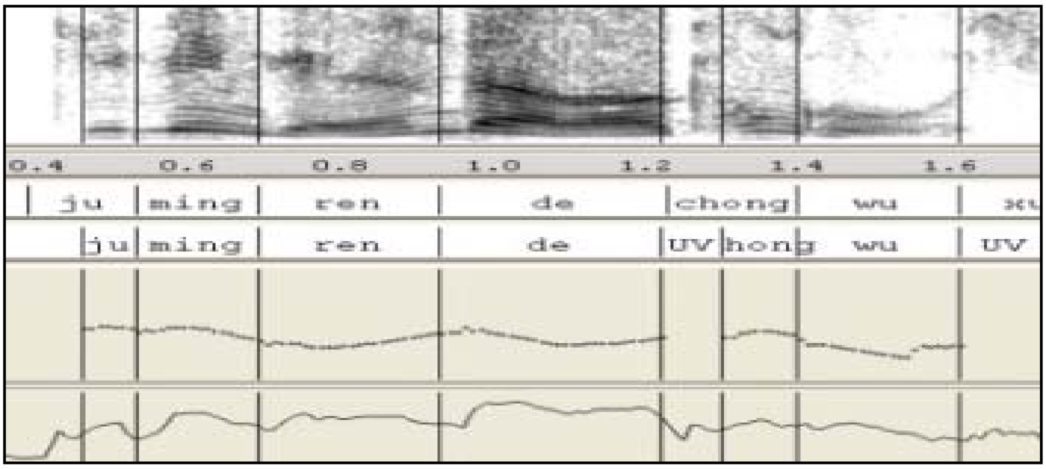

Figure 3. A typical segmentation results of the tokenization front-end (from top to bottom panel: spectrum, syllable and sub-syllable segmentations, log-pitch and log-energy contours).

For each voiced segment, six dimensional prosodic features are extracted including coefficients of 3-order Legendre polynomial function for approximating the log-pitch contour, the log-energy mean and duration of the segment. On the other hand, for each unvoiced segment, only its log-energy mean and duration are utilized.

3.2. LPM training algorithm

To estimate the parameters of the LPM, an unsupervised sequential optimization procedure based on the maximum likelihood criterion is adopted. The training procedure sequentially decodes latent prosody state sequences using Eq. (7) and updates the affecting factors (i.e., tone and prosody state) to optimize the likelihood function in Eq. (6).

In more detail, the sequential optimization training procedure executes the following steps until a convergence has been reached. It is worth noting that each step updates a subset of LPM parameters.

Step 0: Initialization

- Derive the initial affecting factors $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_{t_n}$ of tones by averaging all prosodic feature vector $\mathbf{x}_n$ of a speaker or the whole training data, respectively.
- Cluster and label the prosody state of each segment by vector quantization (VQ) using the residue prosodic feature vector $\mathbf{x}'_n = \mathbf{x}_n - \boldsymbol{\mu}_s - \boldsymbol{\mu}_{t_n}$ and derive the initial prosody state affecting factors $\boldsymbol{\mu}_{q_n}$.
- Derive the initial covariance matrix $\boldsymbol{\Sigma}$.
- Derive the initial prosody state transition probabilities using the statistics of labeled prosody states.

Step 1: Re-Label

- Re-label the prosody state sequence of all utterance using Eq. (7).

Step 2: Re-Estimate

- Update the affecting factors $\boldsymbol{\mu}_s$ of speakers, $\boldsymbol{\mu}_{t_n}$ of tones or $\boldsymbol{\mu}_{q_n}$ of prosody states with all other parameters fixed.
- Update the covariance matrix $\boldsymbol{\Sigma}$ and the prosody state transition probabilities.

Step 3: Iteration

- Repeat step 1 to 2 until the likelihood function Eq. (6) is converged.

4. Experimental Results

4.1. Corpus

To evaluate the proposed LPM approach, two telephone speech corpora were mixed together, one is Mandarin across Taiwan (MAT) [11] released by Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taiwan, and the other is 500-people telephone reading speech corpus (TRSC) [12] released by Chinese Corpus Consortium (CCC), China. There are about 4500 (MAT-2000+MAT-2500) Taiwanese and 500 Mainlander speakers in MAT and TRSC, respectively. The mixed corpus is randomly divided into a training, a development and a test set. The detail of speaker and utterance information is listed in Table. 1. The evaluation is executed utterance by utterance and the average length of an utterance is about 5 seconds.

Table 1. Detail information of the MAT ad TRSC corpora including number of speakers and utterances.

	Training		Development		Test	
	spk	utt	spk	utt	spk	utt
MAT	3936	67633	3742	20192	238	2009
TRSC	409	43340	120	12594	20	2042

4.2. LPM training results

For all following LPM experiments, the number of prosody states was empirically set to 11 (8 for voiced, 3 for unvoiced states) and there are 5 different tones in Mandarin.

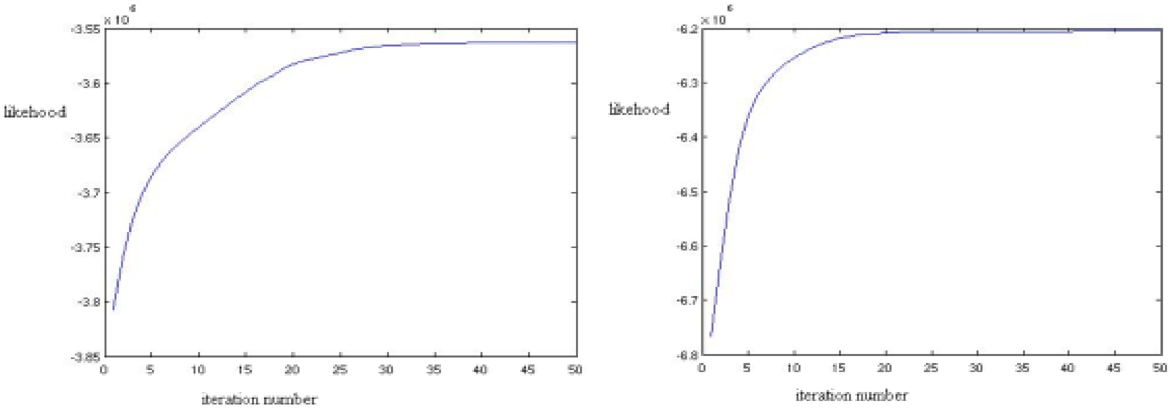

Figure 4. The learning curves of the LPMs training on MAT and TRSC training sets (left: MAT, right: TRSC), respectively.

First of the all, the learning curves of the LPMs were examined. Fig. 4 shows the likelihood functions on the MAT and TRSC training sets, respectively, along with the number of training iterations. It could be found from the figure that LPMs converged quickly, especially for the TRSC set.

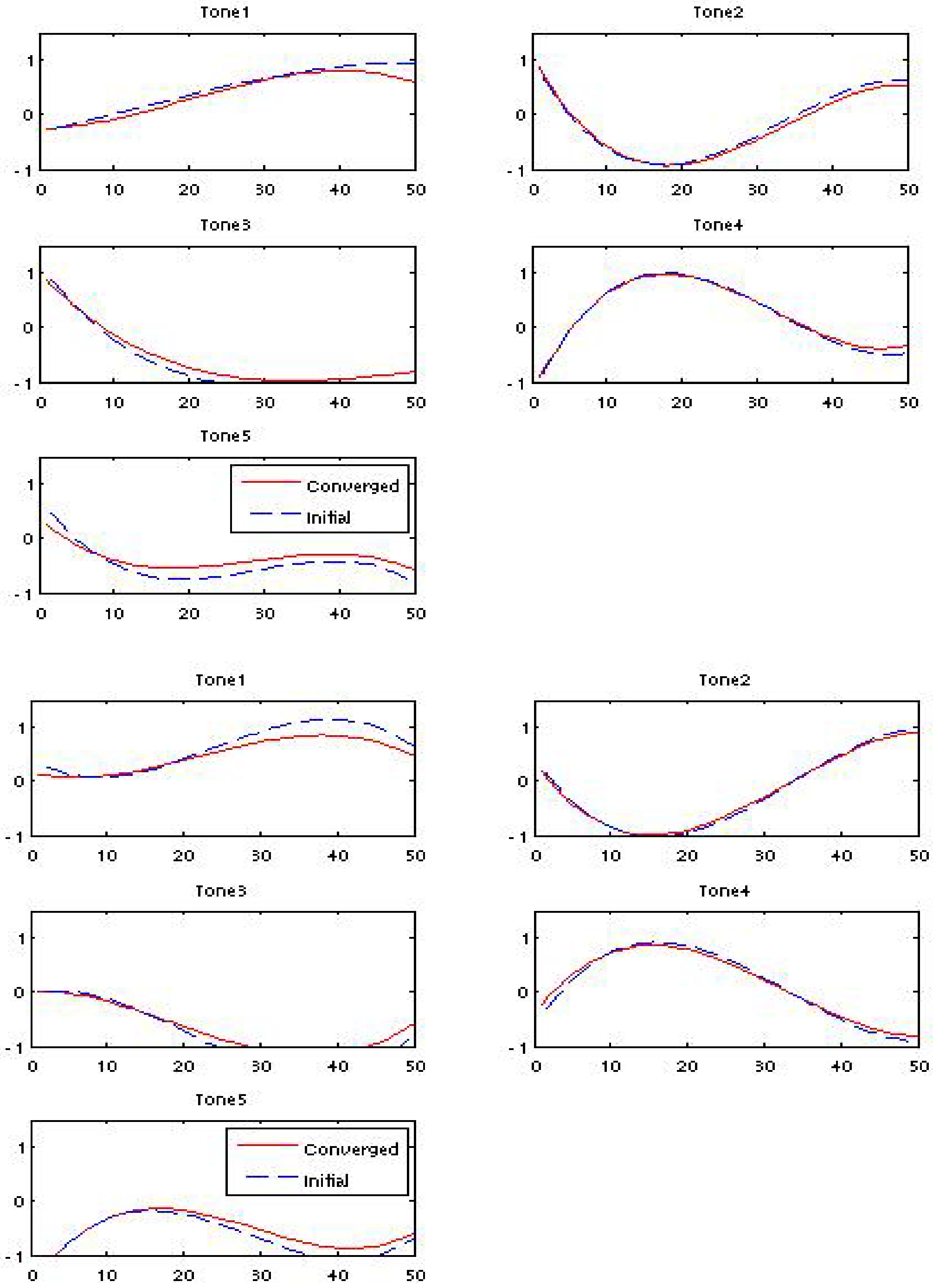

Figure 5. The learned tone affecting patterns on MAT and TRSC corpora (top 5 panels: MAT, bottom 5 panels: TRSC), respectively.

After LPM training was converged, the learned 5 tone affecting patterns of Taiwanese and Mainlanders' Mandarin, respectively, were drawn in Fig. 5. It is found that the major tone differences between Taiwan and Mainland China is the pattern of tone 3 and 5. This is consistent with common linguistic knowledge [10].

These results suggest that LPMs could automatically learn the accent-specific characteristics of Taiwanese and Mainlanders' Mandarin. We therefore expect that LPM-LM-based approach could be successfully used to discriminate these two Mandarin accents.

4.3. Acoustic and Phonotactic baselines

To set up a reference baseline, two popular phonotactic and one acoustic approaches were first tested including (1) PPR-LM, (2) universal phone recognizer (UPR)-LM and (3) shifted delta cepstral (SDC)/Gaussian mixture model (GMM).

For PPR-LM and UPR-LM, 39-dimensional mel-frequency cesptrum coefficient (MFCC) feature vectors were utilized to train the front-end phone recognizers. There are in total 50 phonemes in Mandarin for PPR-LM. But for UPR-LM, the number of phonemes is extended to 63 to reflect the major pronunciation differences (retroflex and nasal-endings sounds) between Mainlander's and Taiwanese Mandarin. All MFCCs were pre-processed by cepstral normalization (CN) to partially compensate the channel and database mismatch. Beside, tri-gram LM backbends were adopted for both PPR-LM and UPR-LM. Moreover, the parameters of SDC were empirically set to 7-3-3-7 and the number of mixtures in GMMs was 512.

Table 2. Experimental results of the individual acoustic, phonotactic and prosodic approaches and their fusion on a mixed TRSC and MAT database.

Approach	Error (%)	System Fusion	Error (%)
(1): PPR-LM	24.88	(5): (1)+(2)	21.84
(2): UPR-LM	23.79	(6): (1)+(3)	22.53
(3): SDC-GMM	29.11	(7): (1)+(2)+(3)	**20.68**
(4): LPM-LM	31.34	(8): (7)+(4)	**16.18**

Table 2 shows the performances of the individual systems and their fusion results. The fusion was done using a softmax-output multi-layer perceptual (MLP) and trained with the development sets. From Table 2, it is found that (1) PPRLM and UPRLM worked better than SDC/GMM and (2) the best performance, 20.68% error rate, was achieved by the fusion of the PPR-LM, UPR-LM and SDC/GMM systems.

4.4. Prosodic approach

The proposed LPM-LM approach was then evaluated. In training phase, the correct tone tags were given but in testing phase MLP-based tone recognizers are adopted to provide estimated tone tags online [7].

Table 2 shows the performances of the proposed LPM-LM and the fusion of LPM-LM with the acoustic and phonotactic baseline. The fusion was also done using the same softmax-output MLP and trained with the development sets. Different from acoustic feature, the prosodic feature extracts another characteristic (example: tone). From Table 2, it is found that LPM-LM worked compatible with the SDC/GMM but is worse than the acoustic and phonotactic baseline. It was caused by just using prosodic feature rather than strong acoustic feature. However, the fusion of LPM-LM and the acoustic and phonotactic baseline could further reduce the error rate from 20.68% to 16.18%. This result may suggest the complementary of those methods.

5. Conclusions

In this paper, a LPM-LM-based approach is proposed to identify two Mandarin accent types spoken by native speakers in Mainland China and Taiwan. Experimental results on a mixed TRSC and MAT database showed that fusion of the proposed LPM-LM and a SDC/GMM+PPR-LM+UPR-LM baseline system could further reduced the average accent identification error rate from 20.7% to 16.2%. Therefore, the proposed LPM method is a promising approach.

6. Acknowledgement

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 96-2221-E-027-100-MY2 and is a partial result of Project 8353C41220 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

References

[1] Language Recognition Evaluation, National Institute of Standards and Technology, http://www.itl.nist.gov/iad/mig/tests/lre/.

[2] Jean-Luc Rouas, "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 1904-1911, Aug. 2007.

[3] Bin Ma, Donglai Zhu, and Rong Tong, "Chinese Dialect Identification Using Tone Features Based on Pitch Flux," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Toulouse, France, May 2006, pp. I-I.

[4] Chi-Yueh Lin and Hsiao-Chuan Wang, "Language Identification Using Pitch Contour Information in the Ergodic Markov Model," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Toulouse, France, May 2006, pp. I-I.

[5] Obuchi, Y. and Sato, N, "Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, Philadelphia, Mar. 2005, pp. 569-572.

[6] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 17, pp. 2095-2103, Sept. 2007.

[7] Chen-Yu Chiang, Xiao-Dong Wang, Yuan-Fu Liao, Yih-Ru Wang, Sin-Horng Chen, and Keikichi Hirose , "Latent Prosody Model of Continuous Mandarin Speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Hawaii, Apr. 2007, pp. IV-625-IV-628.

[8] Chiu-yu Tseng, Shao-huang Pin, Yehlin Lee, Hsin-min Wang, and Yong-cheng Chen, “Fluent speech prosody: Framework and modeling,” *Speech Comminication*, vol. 46:3-4, pp. 284-309, Mar. 2005.

[9] Sin-Horng Chen, Wen-Hsing Lai, and Yih-Ru Wang, ” A statistics-based pitch contour model for Mandarin speech,” *Journal of the Acoustical Society of America*, 117 (2), pp. 908-925, Feb. 2005.

[10] Chin-Chin Tseng, “Prosodic Properties of Intonation in Two Major Varieties of Mandarin Chinese: Mainland China vs. Taiwan,” in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, Beijing, China, Mar. 2004, pp. 28-31.

[11] Hsiao-Chuan Wang, Frank Seide, Chiu-Yu Tseng, Lin-Shan Lee, “MAT-2000 - Design, Collection, and Validation of a Mandarin 2000-Speaker Telephone Speech Database”, in *ICSLP 2000*, Beijing, China, Oct. 2000, pp. 460-463.

[12] 500-People TRSC (Telephone Read Speech, Corpus), Chinese Corpus Consortium, China, http://www.d-ear.com/CCC/corpora/2003-TRSC.pdf, 2003.