INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Pronunciation Evaluation of Non-Native Mandarin Tone by Using Multi-level Confidence Measures

Ju Lin, Yanlu Xie, Jinsong Zhang

College of Information Sciences, Beijing Language and Culture University, Beijing 100083, China
linjucs@163.com, {xieyanlu, jinsong.zhang}@blcu.edu.cn

Abstract

Automatic evaluation of tonal production plays an important role in a tonal language Computer-Assisted Pronunciation Training (CAPT) system. In this paper, we propose an automatic evaluation method for non-native Mandarin tones. The method applied multi-level confidence measures generated from Deep Neural Network (DNN). The confidence measures consisted of Log Posterior Ratios (LPR), Average Frame-level Log Posteriors (AFLP) and Segment-level Log Posteriors (SLP). The LPR was calculated between the correct tone model and competing tone models. The AFLP and LPR were obtained from frame-level scores. And the SLP was directly derived from segment-level scores. The multi-level confidence measures were modeled with a support vector machine (SVM) classifier. For comparison, three experiments were conducted according to different features: AFLP+LPR, SLP only and AFLP+LPR+SLP. The experimental results showed that the performance of the system which used multi-level confidence measures was the best, achieving a FRR of 5.63% and a DA of 82.45%, which demonstrated the efficiency of the proposed method.

Index Terms: CAPT, mandarin tone, automated evaluation, DNN, multi-level confidence measures, SVM

1. Introduction

With the development of globalization, more and more people are willing or required to learn a second language (L2). Computer-Assisted Pronunciation Training (CAPT) is an invaluable resource for L2 learners, as it provides much flexibility in scheduling at low costs. In the past few years, much progress has been made in the development of CAPT system, in which pronunciation evaluation plays an important role. Many studies have been conducted in evaluating Mandarin pronunciation and most of them have focused on segmental goodness [1-5]. However, it is known that Mandarin is a tonal language, in which tone plays a rather important role in distinguishing lexical meanings and is greatly relevant to speech prosody. Automatic evaluation of tonal production is thus necessary for a Mandarin CAPT system.

Most previous studies on tone evaluation have focused on tones over monosyllables. For example, [6] used log-posterior probability as a measure of goodness of tone pronunciation, and achieved an accuracy of 90% and a false acceptance rate of 4%. [7] employed a similar method but optimized the performance of fundamental frequency extraction and the boundary of tone-patterns. The correct rate for tone scoring has been reported to increase from 62% to 83.3%. [8] proposed a method of the normalization of cumulative distribution function that could remove the differences of F0 features between speakers, and made the Cross-Correlation between human experts and automatic tone error detection system is close to 0.79.

While the above-mentioned studies have achieved promising results for monosyllables, only few studies have been done on tone evaluation in connected speech. [9] used the Context Depended Tone Model (CDTM) and Kullback-Leibler Divergence (KLD) between models to detect tone errors. They got the equal error rate at 6.7%. [10] applied the mixed models that combined both short and long time fundamental frequency features, and got an average score error rate at 24.9%. In addition, other methods based on the Goodness of Pronunciation (GOP) [11] were also proposed for tone evaluation. Example, [12] employed a SVM classifier to model the Goodness of Tone (GOT) features extended form GOP. By computing a vector of confidence scores for all possible lexical tones without consideration of the threshold choice, they achieved a better performance. However, lexical tones show complex variations in connected speech, especially in non-native speech. More tone-related information is required for robust tone evaluation. Further study is thus in great need on tone evaluation in connected speech.

In this paper, we focus on tone evaluation in connected Mandarin speech and proposed an automatic evaluation method for non-native Mandarin tones. The method applied multi-level confidence measures generated from Deep Neural Network (DNN). We first trained a frame-level feature based tone model (FLTM) and a segment-level feature based tone model (SLTM), and then used them to get the multi-level scores. The Average Frame-level Log Posteriors (AFLP) and Log Posterior Ratios (LPR) were obtained from frame-level scores. The Segment-level Log Posteriors (SLP) was directly derived from segment-level scores. These confidence measures were used as the input features of a support vector machine classifier (SVM).

The paper is organized as follows: In Section 2, a description of tone modeling will be presented. Section 3 presents multi-level confidence measures. It is followed by an overview of tone evaluation in Section 4. Section 5 gives experiments and results. The paper is concluded with our directions for future work in Section 6.

2. Tone Modeling

As is well known, Mandarin is a tonal language. Each syllable consists of an optional Initial (onset), an obligatory Final (rhyme) and a lexical tone. For example, there are four full lexical tones in Standard Chinese: Tone 1 (high), Tone 2 (rising), Tone 3 (low then rising), and Tone 4 (high then falling), as well as Neutral Tone (lack of lexically specified

http://dx.doi.org/10.21437/Interspeech.2016-1162

tone). To obtain the confidence measures, we first trained two tone models based on DNN.

2.1. Frame-level feature based tone model (FLTM)

As discussed above, there are five lexical tones in Mandarin speech, but there is no tone pattern for silence and unvoiced parts within a syllable. Therefore, we trained a FLTM based on DNN-HMM with six targets: five tones and a no-tone target. As the articulatory features (AF) were found to be useful in Mandarin tone recognition [13], we thus incorporated the AF into the FLTM. Phonological studies have suggested that both the Initials and the Finals can be further divided into a series of detailed categories based on the articulatory movements such as the manner of articulation and place of articulation etc. The articulatory categories which we adopted were consistent with [13] except that we added a SIL symbol for parts of silence, so that the number of articulatory categories was 20 (i.e., 4 for Initials + 15 for Finals + 1 for silence).

The input features of the FLTM consist of two parts. First, the 13-dimensional MFCC and F0 features were spliced in time by taking a context size of 7 frames composing of 3 preceding frames, current frame and 3 succeeding frames, followed by de-correlation and dimensionality reduction to 40 using Linear Discriminant Analysis (LDA) [14]. The obtained features were further de-correlated with the Maximum Likelihood Linear Transform (MLLT) method [15], which was followed by speaker normalization using feature-space Maximum Likelihood Linear Regression (fMLLR) [16]. Second, the posterior probabilities of the articulatory categories were obtained from an articulatory DNN classifier.

2.2. Segment-level feature based tone model (SLTM)

For each syllable, the curve fitting parameters of the F0 contour, duration, average energy and average F0 value were used for SLTM. F0 was by using RAPT [17] as implemented in ESPS's get_f0 (parameters: wind_dur=0.01, min_f0=60, max_f0=650) and normalized to have mean 0 and variance 1 within voiced regions on a per-speaker basis. The F0 curves were fitted with two-order linear regression: $f(t) = at^2 + bt + c$ and the fitting parameters {a, b, c} were used for tone recognition. With consideration of possible tonal co-articulation, features of neighboring tones were also included for the modeling of the current tone. All the obtained features were used as input of the DNN-based SLTM. All the prosodic features are listed in Table 1.

Table 1. *The features of SLTM.*

	Features	Number of Features
1	Curve fitting parameters of the F0 contour	3
2	Duration of the current syllable	1
3	Mean F0 of current syllable	1
4	Mean Energy of current syllable	1
5	All the above features of neighboring syllables	12

3. Multi-level Confidence Measures

In this module, we adopted the AFLP and LPR calculation following [18]. Although the final results of AFLP and LPR were at tone phone level, we regarded the AFLP+LPR and SLP as different levels of confidence measure because AFLP and LPR were converted from frame-level scores, and SLP was directly derived from segment-level scores.

3.1. Averaged Frame-Level Posteriors

In the DNN-HMM based FLTM training, multi-layer neural networks were trained to provide posterior probability estimates for the HMM states, namely, sub-phones ("senones"). We can directly use the posterior probability of "senone" given the parameter observations, instead of converting it to back to HMM's emission likelihood. Here the tone phone posterior is approximated by

$$\begin{aligned} AFLP(p) &= logp\left(p|\boldsymbol{o};t_s,t_e\right) \\ &\approx \frac{1}{t_e - t_s + 1}\sum_{t=t_s}^{t_e} logp\left(p|\boldsymbol{o}_t\right) \end{aligned} \quad (1)$$

$$p\left(p|\boldsymbol{o}_t\right) = \sum_{s\in p} p(s \mid \boldsymbol{o}_t) \quad (2)$$

$p(s|\boldsymbol{o_t})$ is the output of the last "softmax" layer of our DNN model, s is the senone label of the frame t derived from force alignment with the given canonical phone p, $\{s|s \in p\}$ is the set of all senones corresponding to p, i.e., the states of those triphones (HMM models) whose current tone phone is p. $\boldsymbol{o}_t$ is the parameter input observations of the frame t; and t_s or t_e is the start or end frame index of tone phone p, respectively.

As there would be inevitable mismatch between the training utterances produced by native speakers and the testing utterances by non-native speakers, this will lead to inaccurate forced-alignment results at senone level. To address this problem, in the Eq. (2), all possible senones were considered in tone phone posterior to robustly evaluate tonal production, i.e., the senone boundaries are relaxed to phone boundaries.

3.2. Log Posterior Ratios

The Log Posterior Ratio (LPR) between tone phone p_j and p_i is defined as:

$$LPR\left(p_j|p_i\right) = logp\left(p_j|\boldsymbol{o};t_s,t_e\right) - logp\left(p_i|\boldsymbol{o};t_s,t_e\right) \quad (3)$$

where $logp(p|\boldsymbol{o})$ was calculated by Eq. (1).

3.3. Segment-level Log Posteriors

We obtained the boundary information of Initials and Finals of each syllable determined by forced alignment, and then extracted segment-level features which were described in Table 1 of the Final part of the syllable. The corresponding posterior probability of tone was obtained from the trained SLTM. We directly used the posterior probability of tone from the output of the softmax layer of SLTM. It was obtained with

$$SLP\left(t_k\right) = logp\left(t_k|\boldsymbol{o}\right) \quad (4)$$

where t_k is label of tone, k= {1, 2, 3, 4}, and $\boldsymbol{o}$ represents parameters of the syllable.

4. The Framework of the Evaluation of Mandarin Tone

Studies have shown that the classifier-based approaches achieved better performance than GOP-based methods in mispronunciation detection [18, 19]. Therefore, in the present study, SVM classifier was used for mandarin tone evaluation. Different from GOP based method, an extra supervised process is required. The input features for supervised training were segmental features derived from multi-level confidence measures. As illustrated in Figure 1, these features of each segment $\boldsymbol{o_i}$ were calculated from frame posterior matrix, outputs of FLTM, with its start and end frame indexes (t_s, t_e) obtained via forced alignment, as well as segment-level posterior obtained from SLTM. Thus, the segmental feature vector of canonical tone phone p_i is defined as follows:

$$[AFLP(p_1), AFLP(p_2), \dots, AFLP(p_M), LPR(p_1|p_i), LPR(p_2|p_i) \\ \dots, LPR(p_M|p_i), SLP(p_1), SLP(p_2), \dots, SLP(p_M)]^T \quad (5)$$

where M is the total number of all tone phones, and M = 4.

Each tone was manually labeled as correct or incorrect tone pronunciation, which was used as the output labels. SVM classifier was trained based on the input and output features of training data. For SVM classifier training, four individual 2-class SVM classifiers were trained with the LibSVM toolkit [20] where linear kernel was used. Note that each tone had one classifier and we only considered tones 1-4 in this study.

5. Experiments

5.1. Corpora

5.1.1. Native Mandarin Corpus: Chinese National Hi-Tech Project 863

Both the FLTM and SLTM were trained from the Chinese National Hi-Tech Project 863 corpus [21]. Only data produced by female speakers were used, which consist of 48373 utterances from 83 speakers, approximately 56 hours, where each utterance length is about 12 syllables. Among all the female speakers, 74 were used for training and the remaining nine for testing. The training set and the testing set did not have any overlap at speaker-level and utterance-level.

5.1.2. Non-Native Mandarin Corpus: BLCU inter-Chinese speech corpus

We collected a large scale of Chinese L2 speech database, which can be referred to as BLCU inter-Chinese speech corpus [22]. In this paper, we only focused on the continuous speech produced by Japanese learners. Each Japanese learner was asked to read 301 utterances in Mandarin. The average length per-utterance is 7 syllables. All recordings were made in a sound-proofing speech lab, with the sampling rate of 16 kHz encoded in 16-bit pulse-code modulation (PCM). All sound files were manually annotated by trained phoneticians. Three annotators were asked to annotate correct tone label or incorrect tone label according to perceptual listening tasks for each Japanese learner. The final label of the tone was determined via a voting mechanism. Data from two Japanese learners were used as the experimental data, which included 602 utterances about 3942 full lexical tone samples. Among them, 80% were used as the training data for SVM classifiers and the rest as the test set (see Table 2 for more details).

Table 2. *The detail statistics of non-native data.*

	Total number	*Correct*	*Incorrect*
Tone 1	828	588	240 (29%)
Tone 2	862	552	310 (36%)
Tone 3	950	722	228 (24%)
Tone 4	1302	1042	260 (20%)
Total	3942	2904	1038 (26%)

5.2. Tone Recognition Setup

5.2.1. FLTM setup

The articulatory features combining with the MFCC, F0 features were used as the input features of the DNN-HMM based FLTM. For each frame, a total of 60 dimensions (including 20-dimensional articulatory features and 40-dimensional MFCC and F0 features) features were used. The FLTM was trained to distinguish the six targets according to 11 frames features including five preceding frames, the current frame and five following frames. Therefore, the input size of the FLTM was 660. The tone model was initialized with stacked restricted Boltzmann machines (RBMs) that were pre-trained in a greedy layerwise fashion [23]. After pre-training, all weights and bias were discriminatively trained by optimizing the cross entropy between the target (correspond to context-dependent HMM states, there were about 204 states in our experiment) probability and actual output of softmax output with Back-Propagation (BP) algorithm [24]. The details of our DNN-HMM setup are as follows:

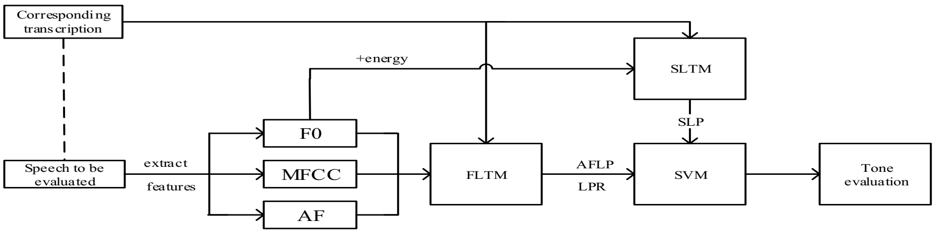

Figure 1: *Schematic diagram of tone evaluation by using multi-level confidence measures*

- A 660-unit input layer.
- 6 hidden layers, each layer consists of 2048 sigmod units.
- An output layer consists of 204 softmax units

5.2.2. SLTM model setup

We compared different network topologies to tune the highest tone classification accuracy. The final full network topology consists of:

- An 18-unit input layer;
- 4 hidden layers: each layer consists of 2048 rectified linear units (ReLUs) [25];
- An output layer consists of 5 softmax units.

The network was trained for 100 epochs using SGD with a mini-batch size of 128, 20% dropout [26] in the input layer, 40% dropout in the hidden layers, a cross-entropy objective.

5.3. Results

5.3.1. The results of Tone recognition

Tone model is the important part of the tone evaluation. Table 3 illustrates the tone recognition results of the FLTM and SLTM in native speech. In FLTM, the performance was significantly improved with a relative error reduction of about 10.1% after incorporating AF.

Table 3. *The Tone Error Rate (TER) of different tone model in native data.*

		TER (%)
FLTM	without AF	9.73
	with AF	8.75
SLTM	16.18	

5.3.2. Tone Evaluation metrics

Three metrics were used to inspect the evaluation performance:

- False Rejection Rate (FRR): The percentage of correctly pronounced phones that are erroneously rejected as mispronounced;
- False Acceptance Rate (FAR): The percentage of mispronounced phones that are erroneously accepted as correct;
- Diagnostic Accuracy (DA): The percentage of detected phones that are correctly recognized, i.e. the detection result is consistent with the human annotations.

5.3.3. The results of tone evaluation

Three experimental systems corresponded to different kinds of confident measures designed in this paper. System 1 adopted the AFLP and LPR measures which were obtained from FLTM. System 2 only used the SLP measure which was obtained from SLTM. System 3 used the multi-level confidence measures including AFLP, LPR and SLP.

We first used the Receiver Operating Characteristic (ROC) metric to compare the performance of different systems. The ROC curve formulates the relationship between true positive rate (TPR) on Y-axis and false positive rate (FPR) on the X-axis. It means the top left corner of the plot is the ideal point (TPR=1, FPR=0). As shown in Figure 2, the system using multi-level confidence measures (Red line) achieved the best performance of each tone, suggesting that our proposed method was efficient and may take full advantage of tone-related information at different levels.

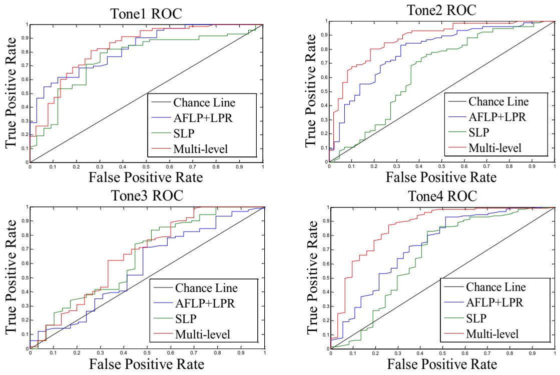

Figure 2 : *ROC curves of evaluations on held-out test set for individual tones*

Considering the purpose of CAPT, DA and FRR are more important in measuring the detection performance than FAR [2, 27]. The evaluation models and the decision threshold were optimized by aiming at maximizing DA. Table 4 further illustrates three metrics of individual tones of system 3, achieving a FRR of 5.63%, a FAR of 49.2% and a DA of 82.45% as a whole. While aiming to maximize the DA and minimize both error rates (FRR and FAR), there is an intrinsic trade-off between the FRR and FAR. Due to the fact that there were much more tones pronounced correctly than those pronounced incorrectly in the corpus, FRR is more decisive than FAR in calculating DA. This has caused a high FAR, especially for Tone 3.

Table 4. *Three metrics of individual tones of systems3.*

	FRR	FAR	DA
TONE1	11.8%	39.5%	78.3%
TONE2	7.1%	38.8%	80.0%
TONE3	2.1%	**70%**	81.7%
TONE4	1.5%	48.4%	89.8%
AVERAGE	5.63%	49.20%	82.45%

6. Conclusions

In this paper, we proposed an automatic evaluation method for non-native Mandarin tones. The method applied multi-level confidence measures generated from Deep Neural Network. Results have proved the efficiency of our proposed approach. The Frame-level and Segment-level confidence measures show a good complementarity for tone evaluation on non-native Mandarin tonal production. In the near future, further efforts will be made to improve the system and more data will be used to develop the tone evaluation of the CAPT system.

7. Acknowledgements

This work is supported by the National Nature Science Foundation of China (61175019) and Wutong Innovation Platform of Beijing Language and Culture University (16PT05) and BLCU support project for young researchers program (16YCX163) (supported by "the Fundamental Research Funds for the Central Universities"). The asterisked author is the corresponding author.

8. References

[1] Ke Yan, and Shu Gong, "Pronunciation proficiency evaluation based on discriminatively refined acoustic models," *International Journal of Information Technology and Computer Science*, pp. 17-23, 2011.

[2] Yingming Gao, Yanlu Xie, Wen Cao, and Jinsong Zhang, "A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network," in *INTERSPEECH*, 2015.

[3] Richeng Duan, Jinsong Zhang, Wen Cao, and Yanlu Xie, "A preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners," in *INTERSPEECH*, 2014.

[4] Feng Zhang, Chao Huang, Frank K. Soong, Min Chu, and Renhua Wang, "Automatic mispronunciation detection for Mandarin," in *ICASSP*, 2008.

[5] Jing Zheng, Chao Huang, Min Chu, Frank K. Soong, and Wei-ping Ye, "Generalized segment posterior probability for automatic Mandarin pronunciation evaluation," in *ICASSP*, 2007.

[6] Li Zhang, Chao Huang, Min Chu, Frank Soong, Xianda Zhang, and Yudong Chen, "Automatic detection of tone mispronunciation in Mandarin," in *ISCSLP*, 2006.

[7] Fuping Pan, Qingwei Zhao, Yonghong Yan, "Improvements in Tone pronunciation scoring for strongly accented Mandarin speech," in *ISCSLP*, 2006.

[8] Si Wei, Hai-kun Wang, Qingsheng Liu, and Renhua Wang, "CDF-matching for automatic tone error detection in Mandarin CALL system," in *ICASSP* 2007.

[9] Yanbin Zhang, Min Chu, Chao Huang, Mangui Liang. "Detecting tone errors in continuous mandarin speech," in *ICASSP*, 2008.

[10] Long Zhang, Haifeng Li, Lin Ma, Jianhua Wang, and Wei Zhang, "Mixed models based pronunciation evaluation of Mandarin tone," *Journal of Multimedia*, vol. 8, no. 6, pp. 726-731, 2013.

[11] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[12] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," in *INTERSPEECH*, 2015.

[13] Hao Chao, Zhanlei Yang, and Wenju Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *ICASSP*, 2012.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork, "Pattern classification," in *Wiley*, 2000.

[15] Ramesh Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *ICASSP*, 1998.

[16] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Comp. Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[17] David Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp.518, 1995.

[18] Wenping Hu, Yao Qian, Frank k. Soong, and Yong Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication* vol. 67, pp. 154-166, 2015.

[19] Wei Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-hui Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *ICASSP*, 2016.

[20] Chih-Chung Chang, and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* (TIST) vol. 2, no.3, pp. 27, 2011.

[21] Sheng Gao, Bo Xu, Hong Zhang and Taiyi Huang, "Update of Progress of Sinohear: Advanced Mandarin LVCSR System At NLPR", in *Proc. ICSLP*, 2000.

[22] Wen Cao, Dongning Wang, Jinsong Zhang, and Ziyu Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *INTERSPEECH*, 2010.

[23] Geoffrey. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[24] David. E. Rumelhart, Geoffrey E. Hinton, and Ronald. J. Williams. "Learning representations by back-propagating errors." *Cognitive modeling*, vol. 5, no.3, pp. 1, 1988.

[25] Vinod Nair, and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807-814, 2010.

[26] Geoffrey E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv: 1207.0580*, 2012.

[27] Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, Jinsong Zhang, "Landmark of mandarin nasal codas and its application in pronunciation error detection," in *ICASSP*, 2016.