Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition

Rong Tong, Nancy F. Chen, Bin Ma and Haizhou Li

Institute for Infocomm Research, Singapore

{tongrong,nfychen,mabin,hli}@i2r.a-star.edu.sg

Abstract

Lexical tone is one of the most challenging pronunciation problems in tonal language acquisition. Accurate lexical tone production is especially challenging for people whose native language is not a tonal one. In this paper, we propose Goodness of Tone (GOT), a confidence measure inspired from goodness of pronunciation (GOP) for tone recognition. GOT is a vector representation of the confidence of each lexical tone of the given speech segment. The proposed GOT confidence measure is useful in tone recognition due to the following: 1) Unlike other tonal features such as pitch or fundamental frequency variation, GOT integrates both phonetic and tonal information. 2) GOT exploits competing tonal phones which differ only in tonal label but are the same in phonetic labels as a reference to conduct cohort normalization. 3) GOT is a vector that concatenates confidence scores from all the possible lexical tones, making it easier to characterize error patterns of non-native tonal production.

Index Terms: automatic speech recognition (ASR), human-computer interaction, computational paralinguistics, computer-assisted pronunciation training (CAPT)

10.21437/Interspeech.2015-254

1. Introduction

Computer-assisted language learning (CALL) systems provide a private and self-paced learning environment to language learners. To help the learner understand the problem and improve his pronunciation, CALL systems provide segmental and suprasegmental level feedbacks on learner's speech input. The suprasegmental feedback focuses on the rhythm, stress, and intonation of the speech [1, 2], while the segmental feedback focuses on the pronunciation accuracy of the individual phonetic units [3, 4].

Mispronunciation occurs from both phonetic and prosodic aspects [5]. Lexical tone error is a special type of mispronunciation present in tonal languages such as Mandarin Chinese. Since the meaning of a word changes with the tone in tonal languages, it is essential to model lexical tone errors of non-native speakers, especially for learners whose native language (L1) is not tonal in nature. This work focuses on the task of Mandarin tone recognition on non-native speech.

Given a non-native speech segment, a tone recognition system compares the user's tone production with the standard pronunciation, and gives the user corrective feedback about the actual tone he produced. This type of feedback is more informative than that of error detection, as it tells the user what the specific error is.

Fundamental frequency (F0) is the acoustic correlate of pitch, the primary feature conveying tones. F0 and its derivatives are commonly used to model tone [6]. A F0 smoothing method is proposed in [7] to improve the performance in Mandarin tone recognition. Tone recognition of German learners of Mandarin are studied in [8], where the spectral and prosodic features are used to recognize tones on monosyllables, disyllables and sentences.

In addition to F0 and its derivatives, several measures of voice quality have also been studied, where band energy features were found to be useful in Mandarin tone recognition[9]. Broad context information, such as the intonation boundaries have also been incorporated into tone recognition systems [10, 11].

Various machine learning techniques have been used for tone modeling. SVM-based tone recognition methods are presented in [12] to model F0 related features. A decision tree based method was proposed in [13] for Mandarin tone recognition. Deep neural network (DNN) classifiers are used in [14, 15] for tone classification in Mandarin broadcast news.

Goodness of Pronunciation (GOP) is widely adopted in pronunciation quality assessment [16]. It measures the phone level confidence by comparing the likelihood ratio of the produced phone and the reference transcription. Refinements of Various segment level GOP scores are derived to achieve better performance in pronunciation assessment [17, 18]. Confused phoneme sets are used in GOP calculation to derive better confidence scores [19]. A lattice based GOP measure [20] was shown to perform better in short sentences and is less sensitive to utterance length.

The GOP algorithm relys on pre-defined thresholds to make mispronunciation decisions. Various thresholds (eg, phone level, segment level and word level) need to be derived from the training data. When lacking of enough non-native data, artificial data is generated to obtain reliable thresholds [18].

In this work, we adapt and extend the concept of GOP to Goodness of Tone (GOT) by computing a vector of confidence scores for all the possible lexical tones. The GOT features are based on posterior probabilities of tonal phones. Given a tonal phone and its posterior probability, we normalize the posterior by considering a cohort of competing tonal phones whose phonetic values are the same as the give phone, but the tonal values differ. We demonstrate that GOT is a simple yet effective measure in modeling non-native Mandarin tones, it elevates the difficulty in finding the optimal thresholds.

2. Tone Recognition Models

2.1. Mandarin Syllable Structure and Lexical Tones

Mandarin Chinese is a monosyllabic language, where each character constitutes a single syllable. Each syllable consists of an optional initial (consonant), a final (vowel) and a lexical tone. There are five tones in Mandarin Chinese, Tone 1 (high), Tone 2 (rising), Tone 3 (low then rising), Tone 4 (high then falls) and Tone 5 (neutral or lack of tone). If tone is not considered, there are $\sim$400 distinct syllables in Mandarin Chinese; if we consider tone, there are $\sim$1300 tonal syllables. Neutral Tone 5 has no specific contour; it is equivalent to an unstressed syllable

in English. In this work, we focus on tone 1-4.

In this paper, a syllable is further decomposed into a sequence of phones in acoustic modelling. Phones are modelled together with lexical tones to form tonal phones.

2.2. Tone Recognition using ASR

2.2.1. *Goodness of Pronunciation*

The Goodness of Pronunciation (GOP) [16] is a phone level confidence measure to gauge how a particular phone is pronounced differently compared to a native model. Given phone p, the GOP score can be derived as

$$GOP(p) = \frac{1}{d}\frac{P(O|p)P(p)}{\sum_{q\in Q} P(O|q)P(q)} \quad (1)$$

where O is the acoustic observation; Q is the set of all phones; d is the number of frames the phone p spans; $P(O|p)$ stands for the likelihood of the observation p, it can be obtained by performing forced alignment with the canonical transcription; $\sum_{q\in Q} P(O|q)$ is the likelihood summation of all the phones in the phone inventory, often derived from phone loop decoding. The denominator is often approximated by taking the maximum; i.e., $\max_{q\in Q} P(O|q)$.

2.2.2. *Goodness of Tone (GOT)*

In this work, we propose a posterior probability based tone confidence measure: Goodness of Tone (GOT). While GOP is a scalar value of a confidence score of a given phone normalized by summing over the posterior probability of all possible phones, GOT is a vector of confidence scores of each lexical tone that is normalized over competing tonal phones which are the same phonetically but different in tone.

Given a speech signal, forced alignment (FA) is first performed using its canonical transcription to derive the phone boundaries. The posterior probability of each phone p is $P_{\mathrm{FA}}(p|O)$. A free phone loop (PL) decoding is then performed. The posterior probability of each phone in the phone inventory can be obtained: $P_{\mathrm{PL}}(p|O)$. For each tonal phone p derived from forced alignment, we are interested in those phones that have the same phonetic characteristics, but different tonal information, we denote them as competing phones.

Figure 1 shows and example of the competing tonal phones given the canonical phone for the syllable *DIAN3*. *DIAN3* is a romanized representation of a Chinese word meaning "dot" using Pinyin, consisting of an syllable initial *D*, a syllable final *IAN* and Tone 3. In this example, the syllable *DIAN3* is represented by a phone sequence: *D I3 EA3 N3*, where the final is made up of three tonal phones: *I3 EA3 N3*. For tonal phone $p = I3$, its competing phones are the set $\tilde{p} \in \{I1, I2, I3, I4\}$. The confidence of tone i can be measured by comparing the posterior probability of the competing phone with the posterior derived from the forced alignment:

$$GOT(\tilde{p_i}) = \frac{1}{d_p}\frac{P_{\mathrm{PL}}(\tilde{p_i}|O)}{P_{\mathrm{FA}}(p|O)} \quad (2)$$

where d_p is the duration of the phone p, which is obtained through forced alignment, $i \in \{1, 2, ...m\}$ denotes the tone index, where the total number of lexical tones is m; $m = 4$ in our case of Mandarin.

For a syllable final s (in this example s = {I3, EA3, N3}) defined in forced alignment, the confidence of each tone can be derived by taking the sum over the phone level confidence, for tone i, $s_i = \{I_i, EA_i, N_i\}$:

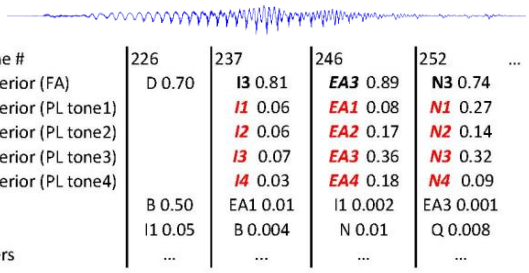

Figure 1: Modeling Goodness of Tone for syllable *DIAN3*. Italicized and bolded tonal phones are the competing cohorts that are phonetically the same but different in tone label when compared to the given phone (obtained through forced alignment).

$$GOT(s_i) = \sum_{\tilde{p_i}\in s_i} GOT(\tilde{p_i}) \quad (3)$$

Thus, for syllable final s, a 4-dimensional GOT feature vector can be derived, where each dimension presents the confidence of each of the four lexical tones.

$$[\, GOT(s_1)\ GOT(s_2)\ GOT(s_3)\ GOT(s_4)\,] \quad (4)$$

We further discuss the advantages of the GOT confidence measure below.

1) Unlike the fundamental frequency estimate (pitch) or fundamental frequency variation (FFV) [21] features, GOT integrates acoustic characteristics from both the phonetic and tonal aspects.

2) Given a tonal phone during force-alignment, GOT only considers cohorts that are the same in phonetic label but differ in tonal label. This procedure can be interpreted as a cohort normalization scheme that is a standard yet effective practice in biometric applications such as speaker verification [22]. The GOT technique is competitive because it chooses cohorts that are as close to the target (forced-aligned tonal phone) as possible.

3) Since GOT is a vector that concatenates confidence scores from the four possible lexical tones, we are able to take into account how each of the four dimensions interact with each other during tone recognition tasks.

2.3. Tone Recognition using Token FFV

In our previous work [23], a token fundamental frequency variation (Token FFV) method was proposed for tone error detection. In this work, we extend the idea of the Token FFV in tone recognition of non-native Mandarin speech.

Fundamental frequency variation is successfully used in various studies [24, 25, 26]. Unlike F0, fundamental frequency variation (FFV) represents pitch variation per frame in vector-form. The derivation of FFV feature is based on the following observation: the rate of F0 change of two adjacent speech frames can be inferred by finding the dilation factor required to optimally align the harmonic spacing in their magnitude frequency spectra [21].

We proposed Token FFV method by using GMM tokenization [27, 28] followed by n-gram language modeling. Given a speech signal, the phonetic boundaries for each syllable can be obtained by performing forced alignment using an ASR system.

With the phonetic boundaries, the corresponding FFV features for each syllable are extracted and the syllables are labeled as Tone 1-4 according to the phonetic transcription.

A GMM universal background model (UBM) is built using all the FFV feature vectors of the training set. For each frame i in the training set, label j is assigned: $j = \arg\max_j P(i|c_j)$, where c_j is the Gaussian mixture component, $j = 1, ..., M$. Thus, each syllable is converted into a GMM index sequence. This GMM index sequence presents the pitch variation of the given syllable.

The GMM index sequences for each of the 4 tone classes are combined and used to derive a tone model using an n-gram language modeling approach. The n-gram language modeling process captures the pitch variation information among n consecutive frames. Compared with the frame-based F0, the proposed Token FFV method captures the pitch variation in relatively longer time spans.

3. Experiments

3.1. Corpora

3.1.1. Native Mandarin Corpus: King-ASR-118

A high performance automatic speech recognition system is a crucial component for Mandarin tone recognition. In this work, a deep neural network based acoustic model is trained from the King-ASR-118 corpus [29]. The King-ASR-118 training set has 326,000 utterances from 975 speakers. The speech in King-ASR-118 are conversations recorded through various type of mobile phones. In order to model the microphone channel effects and reading-style speech of the non-native Mandarin test data, an in-house read speech corpus is used. This corpus is recorded from Mandarin speakers in Beijing and Shanghai in China. Each speaker reads 350 utterances; on average each test utterance is 8 syllables. The read speech training set consists of speech from 450 speakers.

3.1.2. Non-Native Mandarin Corpus: iCALL

The non-native speech corpus used in this study is the iCALL corpus [30, 31]. In this corpus, 305 beginning learners of Mandarin Chinese were asked to read 300 Pinyin prompts. The scripts include 200 short phrases (each phrase has at least two characters) and 100 sentences. Each speaker received a different set of utterances with some overlapped sentences among speakers. The speech was sampled at 16 kHz, encoded in 16 bit pulse-code modulation (PCM), recorded in quiet office rooms. The speech data are manually transcribed in Pinyin through perceptual listening tasks.

A subset of iCALL corpus is split into three portions, training set for acoustic model training, developmental and test sets for tone recognition. These three data sets consists of 237, 30 and 12 speakers respectively. The developmental and test set are selected from the dominating languages of the 3 family groups: the development set consists of speech data of 10 American English speakers from Germanic family, 10 French speakers from Romance family, and 10 Russian speakers from Slavic family. The test set consists of speech of 12 speakers, 4 each from the American English, French and Russian speakers, and there's no overlap between development and test sets.

3.2. Tone Recognition Setup

In tone recognition experiments, we only consider syllables in which the manual phonetic transcription match the reference canonical transcription. For syllables that are incorrectly produced in terms of both phone and tone, we believe that the language learner should focus on getting the phonetic pronunciation correct first since inaccurate phonetic pronunciation affects understanding more than incorrect lexical tones.

Table 1: Number of syllables used in test set

Data	Tone 1	Tone 2	Tone 3	Tone 4
French	872	562	559	804
Russian	812	809	514	985
American English	1038	1010	763	1390
ALL	2722	2381	1836	3179

Table 2: ASR results of the test set

Error (%)	Syllable	Phone	Phone w/o tone
French	56.98	38.05	19.59
Russian	41.99	26.71	14.27
American English	39.32	23.98	13.01
All	46.76	30.11	15.97

Subsets from the developmental and test sets in Section 3.1.2 with no phonetic mistakes were chosen and Table 1 shows the break down details of the test set: the number of syllables for each tone, and the total number of syllables in each group.

3.2.1. Proposed GOT System Implementation Details

The feature vector of the ASR system consists of 13 dimensional MFCC feature in conjunction with 1 dimension of F0, and their derived deltas, acceleration and third-order deltas. The dimension of the feature vector is 56.

The ASR system is trained using the Kaldi toolkit [32]: first, a baseline acoustic model is trained with Maximum Mutual Information (MMI) criterion. Then DNN training is performed using the phone level alignment obtained from the MMI model. There are 5 hidden layers in the DNN models. There are 175 phones and 8,537 tied states. The training portion of the non-native data is incorporated with the native training data only in the DNN training process.

Table 2 reports the automatic speech recognition results of the non-native test set compared against the manual phonetic transcriptions. A syllable loop grammar is used in the decoding process. The recognition errors on both syllable and phone levels are reported. To investigate the importance of tone in Mandarin speech recognition, we report the phone error rate with and without tone. The phone error rate is reduced nearly 50% when tone is not considered. This reveals that mis-recognition of tone is a major challenge for automatic systems.

In this work, the GOT feature vectors are modeled with a support vector machine using one-vs-rest mechanism [33]. For the lexical tone i, an SVM model is trained using GOT vectors derived from segments with lexical tone i in the canonical pronunciation as the positive set and the GOT vectors of the remaining three tones as the negative set.

3.2.2. Baseline Token FFV System Implementation Details

The non-native development set is further separated into two portions with 6:4 ratio. The first portion is used to train two gender dependent UBMs, each has 256 components. The FFV features from the second portion are evaluated on the gender matched UBM to derive the GMM index sequences. The GMM index sequences derived from the same tone class are used to train a 6-gram language model. In the test process, the derived GMM index sequence for each test syllable is evaluated on each of the 4 tone models, the tone model that gives the lowest perplexity is assigned as the tone class.

Table 3: Baseline Tone Recognition with Token FFV

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
French	56.81	41.59	41.58	45.22
Russian	67.44	49.39	32.74	40.09
American English	70.63	47.85	37.36	42.34
All	64.76	46.05	36.78	43.66

Table 4: GOT based tone recognition with general models

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
French	60.89	58.00	53.67	68.53
Russian	74.51	76.27	52.89	84.56
American English	77.46	74.16	62.25	82.44
All	73.23	70.74	56.38	79.34

3.3. Experimental results

3.3.1. Token FFV Baseline

Table 3 reports the tone recognition results of the Token FFV method. For each group, Tone 1 performs the best while Tone 3 is the most challenging. These results are consistent with the human perceptual analyses in [31].

3.3.2. GOT Tone Recognition with General Tone Model

Table 4 reports the tone recognition results of using general tone models. The general tone models are trained with the full development set: regardless of speaker's L1 language background, GOT features of a particular tone from all the speakers are used to train a general tone model.

Compared to the Token FFV tone recognition results shown in Table 3, tone recognition accuracy is improved across all tones. We believe the GOT obtains superior results because it benefits from the richer feature set used in ASR, which containing both phonetic and tonal information, while only tonal information is used in Token FFV. However, similar error patterns are observed in both approaches: Tone 2 and 3 are less accurate than Tone 1 and 4.

3.3.3. GOT Tone Recognition with L1-Dependent Tone Models

It is believed that L1 language background plays an important role in L2 language learning. Due to the native language influence, different tone production errors can be observed in speakers with different L1 background [31]. Figure 2 illustrates the pitch contour of the same syllable: JING3 (J I3 NG3) pronounced by Russian, American English and French speakers. The 3 speakers vary significantly in pitch during the production of tone 3: the Russian speaker has more variations; the American English speaker changes his pitch only at the beginning and end of the syllable final; the French speaker has relatively flat pitch except at the beginning.

To characterize this difference, we build L1 dependent tone

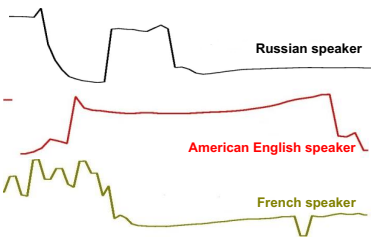

Figure 2: Pitch contour of the same syllable JING3 (J I3 NG3) pronounced by different speakers

Table 5: GOT based tone recognition with L1 Dependent models. The numbers in bracket are the relative improvements (in %) over the general tone model results.

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
French	68.35	63.52	54.38	74.13
Russian	76.11	80.10	51.59	90.86
American English	80.44	83.66	63.77	90.00
All	77.66 (6.05)	78.20 (10.55)	60.29 (6.94)	87.00 (9.65)

US	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**80.44**	4.43	2.7	12.43
Tone 2	1.98	**83.66**	12.77	1.58
Tone 3	1.7	30.67	**63.17**	4.46
Tone 4	5.68	1.37	2.95	**90**

RU	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**76.11**	10.22	3.94	9.61
Tone 2	5.44	**80.1**	11.37	3.09
Tone 3	4.09	39.3	**51.56**	5.06
Tone 4	4.57	1.83	2.64	**90.96**

FR	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**68.35**	11.58	8.94	11.12
Tone 2	14.06	**63.52**	18.86	3.56
Tone 3	4.47	33.81	**54.38**	7.33
Tone 4	13.93	3.11	8.83	**74.13**

ALL	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	**77.66**	7.84	4.43	10.06
Tone 2	6.55	**78.2**	11.88	3.37
Tone 3	3.14	30.17	**60.29**	6.4
Tone 4	6.35	1.99	4.66	**87**

Figure 3: Confusion matrix of L1 dependent tone recognition, US (American English), RU (Rassian), FR (French)

models. For each language group, the GOT features of only that language group are used in tone model training. As a result, we have 3 sets of tone models, one for each L1 language group: French, Russian, American English. During test time, the GOT vectors are evaluated on corresponding tone models belonging to the same L1 language group. The tone recognition results of L1 dependent tone models are reported in Table 5. The numbers in bracket are the relative improvements (in %) over the general tone model results shown in Table 4. The results show that the L1 dependent tone models outperform the general models in every condition.

Figure 3 shows the confusion matrix of the L1 dependent tone recognition for three language groups and all speakers. The GOT based system has the best tone recognition performance for American English speakers, this is consistent with the ASR results shown in Table 2. Tone recognition performance is low for French speakers. One possible reason is that French does not have a significant stress accent (like English), causing French speakers to have a greater tendency of preferring Tone 1 [31].

4. Conclusions

We proposed Goodness of Tone (GOT), a confidence measure for tone recognition. GOT is a vector representation of the confidence of each lexical tone of a given speech segment. The proposed GOT confidence measure is useful in tone recognition because: 1) Unlike tonal features such as fundamental frequency, GOT integrates both phonetic and tonal information. 2) GOT exploits competing tonal phones which differ only in tonal label but are the same in phonetic labels as a reference to conduct cohort normalization. 3) GOT is a vector that concatenates confidence scores from all the possible lexical tones, making it easier to characterize error patterns of non-native tonal production. Our experiment results showed that GOT achieves better tone recognition performance than the baseline Token FFV approach. We also show that exploiting L1 dependent tone model outperforms the general tone models.

5. References

[1] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1989–1999, 2000.

[2] Rong Tong, Boon Pang Lim, Nancy F. Chen, Bin Ma, and Haizhou Li, "Subspace Gaussian mixture model for computer assistant language learning," in *ICASSP*, 2014.

[3] Ann Lee and James Glass, "A comparison-based approach to mispronunciation detection," in *SLT*, 2012.

[4] 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.

[5] Silke M Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, 2012.

[6] Bin Ma, Donglai Zhu, and Rong Tong, "Chinese dialect identification using tone features based on pitch flux," in *ICASSP*, 2006.

[7] Qian Liu, Jinxiang Wang, Mingjiang Wang, Panpan Jiang, Xirui Yang, and Jiayuan Xu, "A pitch smoothing method for Mandarin tone recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 4, 2013.

[8] Hussein Hussein and Hansjg Mixdorff, "Real-time tone recognition in a computer-assisted language learning system for german learners of mandarin," in *Workshop on Speech and Language Processing Tools in Education at COLING-2012*, 2012.

[9] Surendran Dinoj and G-A Levow, "Can voice quality improve Mandarin tone recognition?," in *ICASSP*, 2008.

[10] Siwei Wang and Gina-Anne Levow, "Modeling broad context for tone recognition with conditional random fields," in *INTERSPEECH*, 2011, pp. 2289–2292.

[11] Conggui Liu and Jinxu Tao, "Mandarin tone recognition considering context information," in *Signal Processing, Communication and Computing (ICSPCC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–5.

[12] Lei Wang, Junbo Zhang, Bin Dong, and Yonghong Yan, "A svm based tone recognition for Mandarin multi-syllable words," in *ICASSP*, 2015.

[13] Hsien-Cheng Liao, Jiang-Chun Chen, Sen-Chia Chang, Ying-Hua Guan, and Chin-Hui Lee, "Decision tree based tone modeling with corrective feedbacks for automatic Mandarin tone assessment," in *INTERSPEECH*, 2010.

[14] Neville Ryant, Jia Hong Yuan, and Mark Liberman, "Mandarin tone classification without pitch tracking," in *ICASSP*, 2014.

[15] Xin Lei, Gang Ji, Tim Ng, Jeff Bilmes, and Mari Ostendorf, "DBN-based multi-stream models for Mandarin toneme recognition," in *ICASSP*, 2005.

[16] Silke Maren Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, Cambridge University, 1999.

[17] Jing Zheng, Chao Huang, Mi Chu, Frank K Soong, and Wei-ping Ye, "Generalized segment posterior probability for automatic mandarin pronunciation evaluation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–201.

[18] Ra Kanters, Catia Cucchiarini, and Helmer Strik, "The goodness of pronunciation algorithm : a detailed performance study," in *In SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education*, 2009, pp. 2–5.

[19] Long Zhang, Haifeng Li, and Lin Ma, "Exploit posterior probability algorithm for pronunciation quality evaluation," *Journal of Computational Information Systems*, vol. 8, no. 22, pp. 9251–9258, 2012.

[20] Yin Song, Weiqian Liang, and Runsheng Liu, "Lattice-based gop in automatic pronunciation evaluation," in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*. IEEE, 2010, vol. 3, pp. 598–602.

[21] Kornel Laskowski and Jens Edlund, "A snack implementation and Tcl/Tk interface to the fundamental frequency variation spectrum algorithm," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[22] Douglas E Sturim and Douglas A Reynolds, "Speaker adaptive cohort selection for tnorm in text-independent speaker verification.," in *ICASSP (1)*, 2005, pp. 741–744.

[23] Rong Tong, Boon Pang Lim, Nancy F. Chen, Bin Ma, and Haizhou Li, "Tokenizing fundamental frequency variation for mandarin tone error detection," in *ICASSP*, 2015.

[24] 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*, vol. 117, no. 2, pp. 908–925, 2005.

[25] Hong Xiu Wei, Xin Hao Wang, Hao Wu, Ding Sheng Luo, and Xi Hong Wu, "Exploiting prosodic and lexical features for tone modeling in a conditional random field framework," in *ICASSP*, 2008.

[26] Hussein Hussein, Hansjorg Mixdorff, and Rudiger Hoffmann, "Real-time tone recognition in a computer-assisted language learning system for German learners of Mandarin," in *ICCL*, 2012.

[27] Bin Ma, Donglai Zhu, Rong Tong, and Haizhou Li, "Speaker cluster based GMM tokenization for speaker recognition," in *INTERSPEECH*, 2006.

[28] Pedro A Torres-Carrasquillo, Douglas A Reynolds, and JR Deller Jr, "Language identification using Gaussian mixture model tokenization," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1.

[29] Chinese Mandarin Mobile Speech Recognition Database, "http://www.speechocean.com/en-news/783.html," .

[30] Nancy F. Chen, Vivaek Shivakumar, Mahesh Harikumar, Bin Ma, and Haizhou Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages," in *INTERSPEECH*, 2013, pp. 803–806.

[31] Nancy F. Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li, "iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent," in *Interspeech*, 2015.

[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[33] Rong Tong, Bin Ma, Haizhou Li, and Eng Siong Chng, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 1335–1347, 2009.