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

Context Aware Mispronunciation Detection for Mandarin Pronunciation Training

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

Mispronunciation detection is an important component in a computer-assisted language learning (CALL) system. Many CALL systems only provide pronunciation correctness as the single feedback, which is not very informative for language learners. This paper proposes a context aware multilayer framework for Mandarin mispronunciation detection. The proposed framework incorporates the context information in the detection process and providing phonetic, tonal and syllabic level feedback. In particular, the contribution of this work is twofold: 1) we propose to use a multilayer mispronunciation detection architecture to detect and provide mispronunciation feedback at the phonetic, tonal and syllabic levels. 2) we propose to incorporate the phonetic and tone context information in mispronunciation detection using vector space modelling. Our experiment results show that the proposed framework improves the mispronunciation detection performance in all three levels.

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

1. Introduction

Globalization has increased the needs for people to acquire new languages. Computer-assisted language learning (CALL) systems can assist language learners with automated feedback, allowing them to diagnose, practise and correct language errors at their own pace. Feedback provided by a CALL system can be categorized into two types: (1) Segmental feedback, which focuses on the pronunciation accuracy of the individual phonetic units [1]; (2) Suprasegmental feedback, which focuses on the rhythm, stress, and intonation of the non-native speech [2, 3]. This study focuses more on segmental level Mandarin mispronunciation detection, aiming to provide more informative feedback to language learners.

Pronunciation error patterns in non-native speech can be analyzed to facilitate automatic mispronunciation detection. In [4], linguistic knowledge is utilised to help the mispronunciation modelling for Japanese learners of Mandarin. Data-driven approaches are used to derive pronunciation error patterns automatically in [5, 6], where the error patterns are then used to adapt the native speech trained automatic speech recognition (ASR) system to improve the performance of mispronunciation detection. Phonological rules are extracted from the speech of non-native speakers, and they are used for extending the recognition network for mispronunciation detection [7]. Phone level context [8] has been incorporated in English pronunciation error pattern discovery.

Various features have been explored for mispronunciation detection. Acoustic and prosodic features, such as voicing, articulation rate and duration are used for pronunciation error detection [9]. Tone nucleus can be extracted with vowel landmark detection [10] for Mandarin tone recognition. The landmark of nasal codas in Mandarin are analysed for pronunciation error detection [11]. Word level F0 modeling technique is used in automatic assessment of non-native speech [12]. Acoustic similarity of the non-native speech [13] are extracted for detecting mispronunciation.

The advancement of automatic speech recognition has contributed to that of mispronunciation detection. Many CALL systems improve their performance by adopting Deep Neural Networks (DNN) [14, 15, 16] in acoustic modelling. Goodness of Pronunciation (GOP) and its derivatives have been widely adopted in pronunciation quality assessment [17]. Confused phoneme sets are used in GOP calculation to derive better confidence scores [18]. An aligned GOP method is proposed in [19] for Mandarin mispronunciation detection. Posterior probability vectors can be used to detect phone level mispronunciation for Mandarin [20]. In our previous work, we proposed the use of Goodness of Tone (GOT) [21] for Mandarin tone recognition.

It is found that many mispronunciation detection systems are focused only on the segment itself, while the context information is neglected. However, it is generally agreed that the pronunciation of the individual speech unit is greatly affected by co-articulation. The co-articulation effect is not only observed in phonetic units, it can also occur across phonetic or syllabic units, like in the case of tone sandhi. In this work, we propose to incorporate the context information of different levels in Mandarin mispronunciation detection. We explore the way to incorporate phone and tone level context information by using the phone co-occurrence probability and vector space modelling.

To provide informative feedback to language learners, we propose a multilayer framework which models the different types of pronunciation errors separately. First we focus on the phone level mispronunciation detection without considering the tone, secondly the discriminative tone models are trained for tone level mispronunciation detection, lastly the phone and tone level detection results are combined to derive syllable level feedback.

2. Mandarin Mispronunciation Detection

Mandarin Chinese is a monosyllabic language, where each Chinese character constitutes a single syllable. Hanyu Pinyin is one of the most widely adopted writing systems to romanize Chinese characters. Figure1 shows an example of Chinese Pinyin *SHAN1*. It is the Pinyin form of Chinese character meaning mountain or hill, it consists of an initial (*SH*), final (*AN*) and a tone (*1*). There are five tones in Mandarin Chinese: Tone 1 (high), Tone 2 (rising), Tone 3 (falling then rising), Tone 4

http://dx.doi.org/10.21437/Interspeech.2016-289

Figure 1: An example of Chinese presentation

(high then falls) and Tone 5 (neutral or lack of tone). Since Tone 5 is neutral and has no specific contour, we focus on the mispronunciation of tone 1-4 in this work.

2.1. Mandarin syllable and pronunciation error

In this work, initials and finals are further broken down into smaller acoustic units: phones. Hence syllable *SHAN1* can be represented by a phone string *SH AA1 N1*. A non-native speaker could make three types of mispronunciation at the syllable level:

1) Phone error only: the syllable is phonetically wrong but has correct tone. For example, ***SH*** *AA1 N1* is mispronounced as ***S*** *AA1 N1*.

2) Tone error only: the speaker pronounced the phones correctly, but the tone is produced wrongly. For example, *SH* ***AA1 N1*** is mispronounced as *SH* ***AA4 N4***.

3) Both phonetic and tone errors: the speaker made both phonetic and tone errors at the same time. For example, ***SH AA1 N1*** is mispronounced as ***S AA4 N4***.

2.2. Goodness of Pronunciation (GOP)

Our baseline system adopts the Goodness of Pronunciation (GOP) method for mispronunciation detection. GOP [17] is a phone level confidence measure to gauge how good a particular phone is pronounced compared to a native speech trained model. For segment t, the GOP score for each phone v_i in phone inventory V can be derived as

$$G(i,t) = \frac{1}{d_{v_i,t}} \frac{P(O|v_i)P(v_i)}{\sum_{q \in V} P(O|q)P(q)} \quad (1)$$

where O is the acoustic observation; d is the number of frames the phone v_i spans; $P(O|v_i)$ stands for the likelihood of the observation v_i (it can be obtained by performing forced alignment with the canonical transcription); $\sum_{q \in V} P(O|q)$ is the likelihood summation of all the phones in the phone inventory V, often derived from phone loop decoding.

With GOP scores for each phone, the mispronunciation detection is performed by comparing the GOP score with a set of phone dependent thresholds.

2.3. Multilayer mispronunciation detection

Mandarin tones typically have a longer duration compared to phones. Tonal information is presented throughout the syllable final. For instance, in the syllable ***SH AA1 N1***, tone *1* spans across the two phones: ***AA*** and ***N***.

With the differences between phone and tone in mind, we propose a multilayer mispronunciation detection framework to capture the characteristics of phone and tone separately, and then combine them together for syllable level error detection.

Figure 2 illustrates the proposed multilayer mispronunciation error detection framework. In the first layer, pronunciation of phones are assessed regardless of their tone label, i.e., phones with the same phonetic symbol but different tone are modelled together. In the second layer, regardless of the phone, we consolidate the phones that carry the same tone label in the same model. In the last layer, the outputs of the phone and tone layers are combined to derive syllable level error detection results.

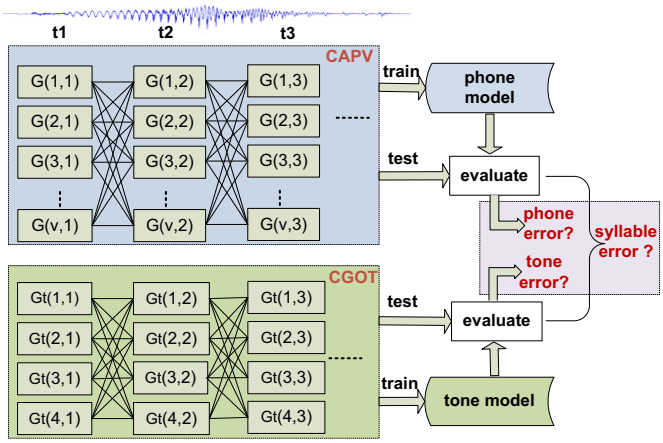

Figure 2: Multilayer mispronunciation detection using Context aware phone vector (CAPV) and Context aware goodness of tone (CGOT)

2.3.1. *Context aware phone vector (CAPV)*

Since it is generally agreed that the pronunciation of the individual speech unit is greatly affected by co-articulation, the mispronunciation detection should benefit from the context information. We propose to use the phone and tone co-occurrence vector to capture context information. This is inspired by the success of the phonotactic feature and use of vector space modeling in language recognition [22, 23, 24]. One advantage of the vector representation is that the interaction of each dimension can be captured in the modelling process.

Without losing generality, Figure 2 shows a syllable which consists of 3 phones. The phone identities and boundaries are obtained from forced alignment. For the phone in segment t_2, its unigram phone vector UV can be composed by concatenating the GOP score of every phone in the phone inventory V, as shown in Eq. 2.

$$UV = [G(1,2)\ G(2,2)\ G(3,2)\ ...\ G(v,2)] \quad (2)$$
$$BV = [B(1,1)\ B(2,1)\ ...\ B(v,1)\ ...\ B(v,v)] \quad (3)$$
$$CAPV = [UV\ BV] \quad (4)$$

To capture the context information, we calculate the co-occurrence of the phone in t_2 and t_1. The bigram posterior probability vector can be obtained by Eq. 3, where

$$B(i,j) = G(i,1) \times G(j,2) \quad (5)$$

it is the product of GOP of phone v_i in this segment t_2 and GOP of phone v_j in the previous segment t_1. It represents the co-occurrence probability of phone v_i being followed by v_j. For a phone at the beginning of a syllable, the end phone of the previous syllable is taken as its previous context.

In this way, a context aware phone vector (CAPV) can be obtained by concatenating the unigram (Eq.3) and bigram (Eq.4) vectors as shown in Eq. 4. For a phone inventory of size n, the dimension of the context aware phone vector will be: $n + n^2$.

2.3.2. *Context aware GOT*

In our previous work, we proposed to use Goodness of Tone (GOT) [25] for Mandarin tone recognition. GOT exploits competing tonal phones which differ only in tonal label but are the same in phonetic labels. In this paper, we extend the GOT to

Table 1: Number of phone and syllable in test set, *w/t* (with tone), *wo/t* (without tone)

Data	syllable		phone	
	correct	**wrong**	**correct**	**wrong**
French *w/t*	2725	2429	9876	4597
French *wo/t*	4727	427	13914	559
Russian *w/t*	3958	1376	11484	2529
Russian *wo/t*	4621	460	13553	460
English *w/t*	4972	639	14692	1100
English *wo/t*	5391	220	15450	342

Context aware GOT (CGOT) following the similar concept as that of the context aware phone vector (CAPV).

The difference between CGOT and CAPV is the content of the phone inventory. In CGOT, the phone inventory consists of only 4 tones. For each segment, the phone inventory only includes those phones with the same phonetic representation but carry different tones. Hence the unigram of a CGOT is the GOP of phones with tone 1 to 4, and the tone context is captured by the co-occurrence probability of the competing tones, as in Eq. 3. Since the tone inventory has size of 4, the vector dimension of the CGOT will be $4 + 4^2 = 20$.

For a given syllable, the CGOT vectors of the same syllable final are summed to form a tone vector for that syllable final. The tone vectors with the same tone labels are used for tone model training.

2.3.3. Mispronunciation detection for syllable

As illustrated in Figure 2, the phone level detection and tone level detection outputs are combined in the last layer to derive syllable level detection results. In the combination process, the same tone label derived from tone level detection is assigned to every phone in the syllable final.

3. Experimental setup

3.1. Corpora

3.1.1. Native Mandarin Corpus

A high performance automatic speech recognition system is crucial for mispronunciation detection. In this work, a deep neural network based acoustic model [25] is trained from the King-ASR-118 mobile speech corpus [26]. The DNN model has 5 hidden layers, 175 tone dependent phones and 8,537 tied states. To capture the characteristics of microphone channel and reading-style speech, an in-house reading speech corpus is used. This corpus is recorded from 450 Mandarin speakers in Beijing and Shanghai in China. Each speaker reads 350 utterances and each utterance has 8 characters on average.

3.1.2. Non-Native Mandarin Corpus: iCALL

The non-native speech corpus used in this study is the iCALL corpus [27]. In this corpus, 305 beginner learners of Mandarin Chinese were asked to read 300 Pinyin prompts.

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

Table 2: Distribution of each tone in test set

Data	Tone 1	Tone 2	Tone 3	Tone 4
French	1834	1043	937	1340
Russian	1721	1397	855	1629
English	1661	1317	1006	1797

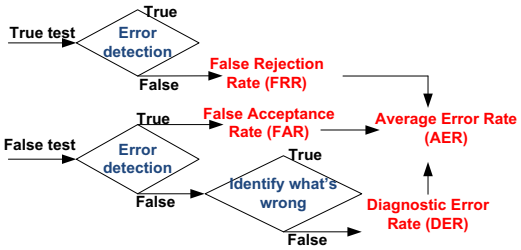

Figure 3: Performance measurements

Table 1 shows the statistics of the test set. The number of syllable and phone of each language and the overall numbers are presented. The determination of correct and wrong pronunciation is based on the manual transcriptions. Both statistics of with tone (*w/t*) and without tone (*wo/t*) are shown. Table 2 shows the distribution of tones in the test sets, the number of syllable for each tone in each language group are reported.

3.2. Performance measure

In this paper, 4 measurements are used to evaluate the mispronunciation detection performance, as illustrated in Figure 3.

1) False Acceptance Rate (FAR): the percentage of mispronounced tests that system failed to detect (also called miss detection);

2) False Rejection Rate (FRR): the percentage of correctly pronounced tests that system erroneously detected as mispronunciation (also called false alarm);

3) Diagnostic error rate (DER): the number of the mispronounced tests that are erroneously diagnosed, divided by total number of correctly detected mispronunciation [6, 7];

4) Average error rate (AER): the average of the previous 3 error rates;

In practical CALL systems, the FRR and DER are considered relatively more important than other measurements. A system with high false rejection rate (FRR) may discourage the learner to continue his study. On the other hand, a system with high detection error (DER) will give user incorrect feedback about the pronunciation. This is not desired as it might mislead the learner to practise with wrong pronunciation.

4. Experiment results

Our baseline system adopts the GOP based method: after forced alignment, the GOP score for each phone is derived and it is compared to a pre-defined threshold. A set of phone dependent thresholds are derived by maximizing the mispronunciation detection performance on the development set.

The results of the GOP baseline system are shown in Table 3. Results of with tone and without tone are reported. The detection error and average error are increased significantly when tone is considered.

4.1. Context aware phone vectors (CAPV)

In this work, the phone inventory size is 43 when tone is not considered. Note that an CAPV vector is a sparse vector, as not all the phones are presented in every decoding hypothesis. For

Table 3: GOP based mispronunciation detection performance, *w/t* (with tone), *wo/t* (without tone)

Test	FAR	FRR	DER	AER
phone wo/t	19.10	11.35	25.34	18.60
phone w/t	13.97	21.72	40.54	25.41
syllable wo/t	13.30	22.78	44.87	26.98
syllable w/t	11.63	37.03	57.01	35.22

Table 4: CAPV based mispronunciation detection, all results do not consider tone (*wo/t*)

Test	FAR	FRR	DER	AER
phone unigram	19.53	11.07	25.20	18.60
phone bigram	21.08	10.22	23.09	18.13
syllable unigram	13.90	22.29	43.55	26.58
syllable bigram	15.20	21.15	41.75	26.03

each phone, an SVM model is trained using the CAPV vectors of that phone from the development set. The phone vector of the test phone is evaluated on the corresponding model. The operation thresholds are decided with development set.

Figure 4 shows the Detection Error Tradeoff (DET) curve which compares the phone level mispronunciation performance of the unigram phone vector and bigram phone vector. The DET curve illustrates the tradeoff between FAR (x-axis) and FRR (y-axis). It is clearly shown that the bigram phone vector gives better performance in phone mispronunciation detection.

Table 4 shows the CAPV performance on the operation thresholds (obtained on the development set). Both phone and syllable level results are reported. Comparing the results in Table 4 with the tone independent results (without tone wo/t) in Table 3, it is clearly shown that the CAPV methods decreases both DER and FRR. This indicates that CAPV have higher ability in detecting mispronunciations than GOP based method.

Between unigram and bigram phone vectors, we see that bigram CAPV offers more consistent reduction of FRR and DER than unigram one. There is a trade off between the FRR and FAR: as FRR decreases, FAR increases. As discussed in Section 3.2, FRR and DER are relative more important in practical applications. This confirms that incorporating context information improves the mispronunciation detection performance.

4.2. Context aware GOT (CGOT)

Table 5 compares the tone error detection accuracy using GOT and CGOT. The overall tone detection accuracy and the break down results for native speakers of each L1 are presented. The CGOT method outperforms the GOT method on the overall tests and also on utterances of the Russian and American English speakers.

Although CGOT improves the overall tone recognition performance for French speakers, there are performance drop for tone 2 and 3 on French utterances. One possible reason is that if the error patterns are inconsistent, suboptimal performance might be observed due to training/testing mismatches [22, 23]. Since French lacks lexical stress, French speakers have more trouble producing lexical tones compared to speakers of other L1 like English [27]. As Tone 2 and Tone 3 are the most confusing Mandarin tonal pairs, we suspect French speakers are more likely to produce inconsistent tonal sequences when Tone 2 and 3 are involved, potentially explaining the dip in performance.

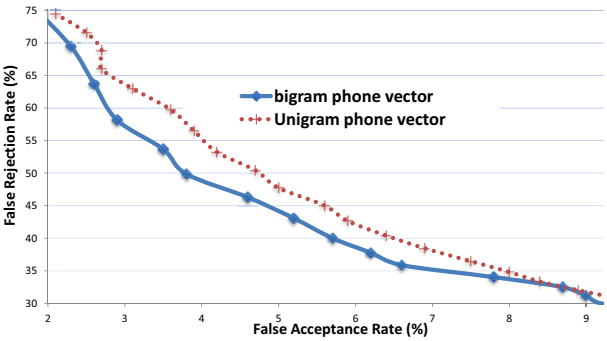

Figure 4: Phone level mispronunciation detection with CAPV

Table 5: Tone detection accuracy of GOT and CGOT

ACC (%)	Tone 1	Tone 2	Tone 3	Tone 4
French GOT	72.68	55.60	52.79	68.80
French CGOT	75.68	53.02	51.40	69.28
Russian GOT	78.15	77.38	47.37	81.61
Russian CGOT	81.23	79.46	51.23	83.52
English GOT	80.25	82.74	58.45	90.60
English CGOT	84.59	83.69	64.31	90.47
All GOT	78.21	74.13	53.76	82.76
All CGOT	81.37	75.23	55.53	83.39

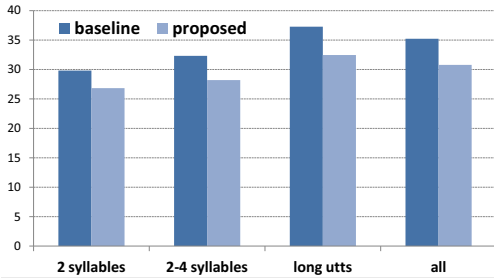

Figure 5: Average error rate (AER) for utterances of different lengths

4.3. Combining CAPV and CGOT

The output of the phone and tone level detection results are combined for syllable level detection. All the final phones in the syllable are assigned the same tone, which is obtained from tone detection. Figure 4 compares the average error rate (AER) of the baseline (GOP based method) and proposed system by utterance length (2 syllables, 2-4 syllables and long utterances).

The results show that the proposed multilayer framework improves the mispronunciation detection performance. It is also found that the performance drops with the increasing of the utterance length. This proves our assumption that the co-articulation plays an important role in mispronunciation detection, while incorporating the context information improves the mispronunciation detection.

5. Conclusion

We proposed a multilayer mispronunciation framework for Mandarin mispronunciation detection. The detection is first performed on phonetic level, followed by the tone level, and the output of the two levels are combined in the last level to provide syllable feedback. In both phonetic level and tone level, we proposed the incorporation of context information through vector space modelling. Our experimental results have shown that our proposed method outperforms the conventional GOP-based method for pronunciation error detection. Incorporating context information improves the mispronunciation performance in most cases.

6. References

[1] Silke M Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, vol. 6, 2012.

[2] 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.

[3] Rong Tong, Boon Pang Lim, Nancy F Chen, Bin Ma, and Haizhou Li, "Subspace gaussian mixture model for computer-assisted language learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5347–5351.

[4] Shuang Xu, Jie Jiang, Zhenbiao Chen, and Bo Xu, "Automatic pronunciation error detection based on linguistic knowledge and pronunciation space," in *ICASSP*, 2009, pp. 4841 – 4844.

[5] Alissa M. Harrison, Wing Yiu Lau, Helen Meng, and Lan Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *INTERSPEECH*, 2008, pp. 2787–2790.

[6] Yow-Bang Wang and Lin-shan Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, pp. 564 – 579, 2015.

[7] Alissa M. Harrison, Wai-kit Lo, Xiao-jun Qian, and Helen Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Slate*, 2009, pp. 45–48.

[8] Ann Lee and James Glass, "Context-dependent pronunciation error pattern discovery with limited annotations," in *Interspeech*, 2014, pp. 2877–2881.

[9] Vaishali Patil and Preeti Rao, "Automatic pronunciation assessment for language learners with acoustic-phonetic features," in *24th International Conference on Computational Linguistics*, 2012, p. 17.

[10] Siwei Wang and Gina-Anne Levow, "Mandarin chinese tone nucleus detection with landmarks," in *INTERSPEECH*, 2008, pp. 1101–1104.

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

[12] Xinhao Wang, Keelan Evanini, and Su-Youn Yoon, "Word-level F0 modeling in the automated assessment of non-native read speech," in *Slate*, 2015.

[13] Christos Koniaris, *Perceptually motivated speech recognition and mispronunciation detection*, Ph.D. thesis, KTH Royal Institute of Technology, 2012.

[14] Wenping Hu, Yao Qian, and Frank K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *INTERSPEECH*, 2013, pp. 1886–1890.

[15] Xiaojun Qian, Helen M. Meng, and Frank K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," in *INTERSPEECH*, 2012, pp. 775–778.

[16] Kun Li, Xiaojun Qian, Shiying Kang, Pengfei Liu, and Helen Meng, "Integrating acoustic and state-transition models for free phone recognition in L2 English speech using multi-distribution deep neural networks," in *SLaTE 2015*, 2015.

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

[18] 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.

[19] Changliang Liu, Fuping Pan, Fengpei Ge, Bin Dong, and Yonghong Yan, "Forward optimal measures for automatic mispronunciation detection," in *ISCSLP 2010*, 2010, pp. 80–83.

[20] Jie Jiang and Bo Xu, "Exploring the automatic mispronunciation detection of confusable phones for Mandarin," in *ICASSP*, 2009, pp. 4833–4836.

[21] Rong Tong, Nancy F. Chen, Boon Pang Lim, Bin Ma, and Haizhou Li, "Tokenizing fundamental frequency variation for mandarin tone error detection," in *ICASSP 2015, April 19-24, 2015*. 2015, pp. 5361–5365, IEEE.

[22] Rong Tong, Bin Ma, Haizhou Li, and Chng Eng Siong, "A target-oriented phonotactic front-end for spoken language recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 7, pp. 1335–1347, 2009.

[23] Haizhou Li, Kong Aik Lee, and Bin Ma, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, pp. 1136 – 1159, May 2013.

[24] Haizhou Li, Bin Ma, and Chin-Hui Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, 2007.

[25] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, "Goodness of Tone (GOT) for non-native mandarin tone recognition," in *INTERSPEECH 2015, Dresden Germany, September 6-10, 2015*, 2015, pp. 801–805.

[26] SpeechOcean, "Chinese mandarin mobile speech recogition database," http://www.speechocean.com/en-News/783.html.

[27] 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, Dresden Germany, September 6-10, 2015*, 2015, pp. 324–328.