A Mispronunciation Detection Method of Confusing Vowel Pair for Chinese Students

Guimin Huang*, Qiupu Chen, Hongtao Zhu
School of Computer Science and Information Security, Guilin University of Electronic Technology, China
sendhuang@126.com
agmhuang@guet.edu.cn

Abstract. This paper presents a discriminative features based mispronunciation detection method for confusing vowel pair /i/ vs /ɪ/ that are frequently mispronounced by Chinese learners of English. Firstly, the mean of the 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) feature vector over all the frames of the current phoneme segment is employed as features to characterize the phoneme. Secondly, many specific acoustic features that can effectively capture the crucial properties of the long and short vowels are extracted. Finally, the Support Vector Machine (SVM) classifier is used for discrimination between confusing vowels /i/ and /ɪ/ by using the discriminative features extracted from each phoneme. The experimental results show that, the proposed method can produce higher accuracy than the traditional Automatic Speech Recognition (ASR) based methods. In addition, the combination of spectral features with specific acoustic features can achieve better performance than using individual features.

1. Introduction

Computer Assisted Pronunciation Training (CAPT) systems provide a platform for second language (L2) learners to practice their pronunciation autonomously at any time and wherever they are, and hence receive growing attention in recent years. Mispronunciation detection is one of the most important components of CAPT systems, which can be very helpful to L2 learners by pinpointing their pronunciation errors and giving relevant feedback[1].

There are two main types of researches on mispronunciation detection at segmental level, one is based on Automatic Speech Recognition (ASR) framework[2-3], and the other is based on classifier. In the ASR based mispronunciation detection method, Confidence Measures (CMs) are calculated and then compared with the appropriate thresholds to decide whether the target phoneme was pronounced correctly. The advantage of the ASR based approach is that these CMs can be computed relatively easily and it can detect pronunciation errors using similar procedures for all phonemes. However, for those confusing phonemes with subtle acoustic differences and mispronounced frequently by L2 learners, the acoustic model has a limited discrimination capability. Therefore, a number of studies have been done to detect these typical pronunciation problems of learners by using dedicated classifiers and discriminative Acoustic Phonetic Features (APF). Strik et al.[4] compared four different classifiers (GOP, APF, LDA-APF, LDA-MFCC) for discrimination between the fricative /ʃ/ and the plosive /k/. The experimental results showed the two LDA classifier based methods outperformed other methods. Maqsood et al.[5] considered mispronunciation detection as a binary classification problem, and used SVM to detect complete mispronunciation of 5 Arabic phonemes. Dang et al.[6] applied SVM to detect
common substitution errors mispronounced by Vietnamese learners of English. All these researches show that, for the detection of specific types of pronunciation errors, the classifier based methods can produce better performances than the traditional ASR based methods.

Most Chinese learners of English find it difficult to discriminate between the long vowel /i/ and the short vowel /ɪ/. The short vowel /ɪ/ is often mispronounced as the long vowel /i/ when they speak English. This may partly be explained by the influence of mother tongue that English has two similar vowels: /i/ and /ɪ/, where Mandarin only uses /i/. In addition, due to the lack of knowledge in the place and manner of articulation, learners usually rely more on the duration than other acoustic features to identify and perceive these two confusing vowels. Detecting this kind of typical substitution errors that are frequently mispronounced by learners is relatively difficult, but it can effectively improve the pronunciation level of learners. On the basis of the above analysis, we propose a discriminative features based mispronunciation detection method which combines the spectral features with specific acoustic features to better discriminate the confusing vowels /i/ and /ɪ/ that are problematic for most Chinese learners of English.

In this paper, the discriminative features based mispronunciation detection method is described in Section 2. The corpus used in the experiments are introduced in section 3. Experimental results are shown in section 4, conclusion and future work follow in section 5.

2. Discriminative Features Based Mispronunciation Detection Method

We design the mispronunciation detection method by the following steps: firstly, a MFCC based spectral feature is proposed; secondly, many specific acoustic features like duration, formant, and pitch are extracted. Finally, we treat the mispronunciation detection of confusing pair /i/ vs /ɪ/ as a binary classification problem and these discriminative features extracted from each phoneme are used to train and test the SVM classifier so as to achieve the mispronunciation detection of /i/ and /ɪ/.

2.1. Spectral feature extraction

The spectral features used in our method are derived from MFCC, which takes into account the auditory characteristics of the human ear, and has proven to be a very effective and robust feature in ASR systems. MFCC features can be extracted by the following steps: Firstly the speech signal is passed through a pre-emphasis filter to amplify the high frequencies. Then the signal is segmented into overlapping frames and a Hamming window is applied to each frame. The power spectrum of each frame is generated by using a Fast Fourier Transform and then mapped into the Mel scale using the Mel Filter Bank. Finally the MFCC features can be obtained by applying a Discrete Cosine Transform (DCT) to the logarithm of all outputs of the Mel Filter Bank. The detailed process of MFCC feature extraction from a speech signal is shown in Fig. 1.

![Fig.1 MFCC feature extraction](image)

After applying DCT, we can get 12 MFCC coefficients and one energy feature per frame. These features describe the static properties of speech signal. In order to capture the dynamic characteristics of the speech signal, the first and second-order derivatives are also calculated and finally the 39-dimensional MFCC feature vector per frame is constructed.

After the above process, each phoneme can be represented as a matrix of 39-dimensional MFCC feature vectors, where, the row represents the number of phoneme frames and the column represents the corresponding MFCC feature vector coefficient. In the matrix, the number of rows for different phonemes may be different, depending on the duration of the phoneme. To address the issue of the difference in the number of frames between phonemes, in this study we take the mean of the column
matrix and finally 39 feature values are obtained to characterize each phoneme. We call this new feature Statistical MFCC (SMFCC).

Compared with the original matrix of the 39-dimensional MFCC feature vectors, the proposed SMFCC feature may impose a loss of information in the phoneme, but it will give a global statistical result, more specifically, the mean of the MFCC feature vectors to represent the phoneme and can reduce the feature dimension. For each phoneme, the final outputs of the spectral feature extraction are 39 feature values. These fixed-length values will be used in the next mispronunciation detection of confusing phonemes.

2.2. Acoustic feature extraction

In addition to the MFCC based spectral features, several specific acoustic features like duration, formant, and pitch are also extracted because of their ability to effectively capture the crucial properties of vowels.

Duration: Duration is the most perceptually salient feature for describing the differences in the length of the sound between the long and short vowels. The vowel /i:/ is a longer sound than /ɪ/ and pronouncing it this way can help distinguish between these two vowels. Most Chinese learners of English give primacy to duration over other acoustic features when they discriminate these two confusing vowels.

Formant: Vowel quality can be determined primarily by the first three formants (F1, F2 and F3), which is related to vowel height, degree of backness and lip rounding, respectively. These characteristics make the formant a very effective feature in the distinction of long and short vowels. In our study, in order to obtain a more accurate calculation result, 50% of the relatively steady portion in the middle of the vowel segment is analysed, and their averages are taken as the final measurements of the first three formants.

Pitch: Among the various acoustic features, pitch is also an important parameter for describing the variations in vowel prosody. The mean pitch of the vowel segment is calculated in our study. Considering that Chinese English learners usually cannot pronounce as fluently as native English speakers, we also carry out the vowel duration normalization by multiplying the raw vowel duration with the articulation rate per speaker as follows:

\[
\text{normalized duration} = \frac{\text{raw duration}}{\text{articulation rate}}
\]

where, articulation rate is defined as the number of phonemes divided by total duration of speech without internal pauses[8]. These 5 normalized features will be referred to as Acoustic.

2.3. Classification with SVM

Since these discriminative features extracted from each phoneme capture some critical properties of the confusing vowels /i/ and /ɪ/, they can be used to train different classifiers to classify these two confusing vowels as either a correct or erroneous pronunciation. In this study, we use the trained SVM classifier to detect whether a vowel is mispronounced as its confusing pair.

3. Corpus

The corpus used in our experiments is taken from the L2-ARCTIC corpus\(^1\), which contains four subsets BWC, LXC, NCC and TXHC recorded by four Chinese speakers (two male, two female). These speakers were asked to read phonetically rich sentences from Arctic corpus. There are 150 manually annotated phonetic transcriptions in each subset, they are produced by the Montreal forced-aligner\(^2\) in PRAAT’s TextGrid format\(^3\). Apart from these 150 annotated sentences per subset, we also manually annotated other 50 utterances that included phoneme substitution of /ɪ/ with /i/. In the end, a total of 800 annotated phonetic transcriptions are used in our study.

For the mispronunciation detection of confusing vowel pair /i/ vs /ɪ/, since the number of substitution of /ɪ/ with /i/ is still relatively rare in the actual annotated material, which is not sufficient to train a separate classifier for each phoneme. So, we focus on the effectiveness of these discriminative features

---

\(^1\) https://psi.engr.tamu.edu/l2-arctic-corpus/

\(^2\) https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner/releases

\(^3\) http://www.fon.hum.uva.nl/praat/
in the discrimination between confusing vowels /i/ and /ɪ/ that have been correctly pronounced. Table 1 gives the correctly pronounced number of phonemes that are used to train and test the SVM classifier.

Table 1. The Number Of Phonemes In Training And Testing Set

| Phoneme | Training set | Testing set |
|---------|--------------|-------------|
| /i/     | 850          | 259         |
| /ɪ/     | 850          | 274         |

In our study, the features that are used for discrimination between /i/ and /ɪ/ are divided into two categories, namely spectral features and acoustic features. The spectral features are derived from MFCC. Here, 12 MFCC coefficients, one energy feature and their first and second derivatives are calculated with a 25 ms window size and 10 ms frame shift. The acoustic features including duration, F1, F2, F3 and pitch are automatically extracted using the PRAAT software. Based on these extracted materials, the SVM classifier with a RBF kernel is trained using the LIBSVM tool4. To optimize the classifier performance, 5-fold cross validation is used.

4. Experiment

In the experiment, we consider four outcomes of correctly accepted (CA), correctly rejected (CR), False Accept (FA) and False Reject (FR) that are similar to the GOP based mispronunciation detection method. CA represents the number of pronounced correctly phonemes that were detected as correct. CR represents the number of mispronounced phonemes that were detected as incorrect. FA represents the number of mispronounced phonemes that were falsely detected as correct. FR represents the number of pronounced correctly phonemes that were falsely detected as incorrect.

Then, the performance of mispronunciation detection for confusing pair /i/ vs /ɪ/ can be evaluated by accuracy, which is defined as follows:

\[
\text{accuracy} = \frac{CA + CR}{\text{Total number of phonemes}} \times 100\%
\]  

(2)

To evaluate the effectiveness of the proposed method, we also developed a baseline GOP method. The SVM classifier was firstly trained and tested using the individual feature, then these two types of features were combined. The experimental results are shown in Fig. 2.

According to the experimental results, we can find that for the mispronunciation detection of confusing vowel pair /i/ vs /ɪ/, our method that is based on discriminative features can yield better results than the traditional GOP method. Especially, the proposed SMFCC feature performed very well and obtained an accuracy of 79.64%. The acoustic features alone can’t perform as well as the SMFCC feature, the accuracy is 75.13%, but the promising performance can be obtained by combining these two types of features, and the accuracy increased to 81.26%. The GOP method produced the worst result, with an accuracy of 70.58%.

![Fig.2 The accuracy for different features](https://www.csie.ntu.edu.tw/~cjlin/libsvm/)

4 https://www.csie.ntu.edu.tw/~cjlin/libsvm/
The proposed SMFCC feature has proved its efficiency by producing a higher classification accuracy than the acoustic features. It is capable to effectively characterize the phoneme in spite of a loss of information. Compared with the SMFCC feature, these specific acoustic features extracted in our experiment are more intuitive and can provide more useful feedback for learners in the articulatory terms. It is often these articulation-related acoustic features that are critical in the identification and perception of confusing vowels /i/ and /ɪ/ for most Chinese learners of English. Furthermore, better performance can be obtained by combining the SMFCC feature with the acoustic features, suggesting that these discriminative features capture some crucial properties that are complementary to describe the confusing vowel pair. Above all, the proposed method is very effective in the mispronunciation detection of confusing English vowel pair /i/ vs /ɪ/ that are problematic for most learners.

5. Conclusion And Future Work
In this paper, we propose a method that combines discriminative features with SVM classifier for mispronunciation detection of confusing vowel pair /i/ vs /ɪ/ that are problematic for most Chinese learners of English. A MFCC based spectral feature is proposed. A number of specific acoustic features like duration, the first three formants and pitch are extracted as well. All these discriminative features extracted from each phoneme are then used to train and test the SVM classifier, so as to achieve the mispronunciation detection of confusing pair /i/ vs /ɪ/. Experimental results show that, compared with the traditional GOP based mispronunciation detection method, the proposed method can yield better results and the spectral features can be combined with the specific acoustic features to further increase the mispronunciation detection performance.

In the future, we will explore more discriminative features and research about feature selection approaches such as dimensionality reduction so as to obtain better effect of mispronunciation detection.

Acknowledgments
This work is supported by the National Natural Science Foundation of China (No. 61662012, No. 61662015).

References
[1] N. F. Chen and H. Li, “Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning,” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, 2016, pp. 1-7.
[2] W. Hu, Y. Qian, F. K. Soong, and Y. 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, March 2015.
[3] J. V. Doremalen, C. Cucchiarini, and H. Strik, “Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch,” The Journal of the Acoustical Society of America, vol. 134, pp. 1336-1347, August 2013.
[4] K. P. Truong, A. Neri, C. Cucchiarini, and H. Strik, “Automatic pronunciation error detection: an acoustic-phonetic approach,” in Proc. InSTIL, Venice, 2004, pp. 135-138.
[5] M. Maqsood, H. A. Habib, T. Nawaz, and K. Z. Haider, “A complete mispronunciation detection system for Arabic phonemes using SVM,” International Journal of Computer Science and Network Security (IJCSNS), vol. 16, pp. 30-34, March 2016.
[6] T. D. Dang and K. G. D. Thi, “Automatic detection of common mispronunciations of Vietnamese speakers of English using SVMs,” in 2017 International Conference on System Science and Engineering (ICSSE), Ho Chi Minh City, 2017, pp. 231-234.
[7] F. Han, “Pronunciation Problems of Chinese Learners of English,” ORTESOL Journal, vol. 30, pp. 26-30, January 2013.
[8] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” The Journal of the Acoustical Society of America, vol. 107, pp. 989-999, January 2000.