Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Landmark-Based Pronunciation Error Identification on Chinese Learning

Xuesong Yang[1], *Xiang Kong*[2], *Mark Hasegawa-Johnson*[1], *Yanlu Xie*[3]

[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[3]School of Information Science
Beijing Language and Culture University, Beijing, China
{xyang45, xkong12, jhasegaw}@illinois.edu, xieyanlu@blcu.edu.cn

Abstract

This paper explores a novel approach of identifying pronunciation errors for the second language (L2) learners based on the landmark theory of human speech perception. Earlier works on the selection method of distinctive features and the likelihood-based "goodness of pronunciation" (GOP) measurement have gained progress in several L2 languages, e.g. Dutch and English. However, the improvement of performance is limited due to error-prone automatic speech recognition (ASR) systems and less distinguishable features. Landmark theory posits the existence of quantal nonlinearities in the articulatory-acoustic relationship, and provides a basis of selecting landmark positions that are suitable for identifying pronunciation errors. By leveraging this English acoustic landmark theory, we propose to select Mandarin Chinese salient phonetic landmarks for the Top-16 frequently mispronounced phonemes by Japanese (L1) learners, and extract features at those landmarks including mel-frequency cepstral coefficients (MFCC) and formants. Both cross validation and evaluation are performed for individual phonemes using support vector machine with linear kernel. Experiments illustrate that our landmark-based approaches achieve higher micro-average f1 score significantly than GOP-based methods.

Index Terms: Pronunciation Error Identification, Acoustic Landmarks, Distinctive Features, Second Language Acquisition

1. Introduction

Pronunciation error identification, as an essential technology in computer assisted pronunciation training systems that provide an effective way of enhancing the speaking skills for the second language (L2) learners, attracts considerable attention from research communities of speech signal processing and applied linguistics.

With the advance in automatic speech recognition (ASR) research, solutions that identify pronunciation errors have made great progress recently. These systems typically detect segmental (phone level) mispronunciations from L2 learner's read speech using an ASR decoder, and pinpoint salient pronunciation errors, such as insertions, substitutions, or deletions of specific pronunciation units [1]. More specifically, two types of ASR-based mispronunciation detection techniques have been widely applied. *Rule-based* approach uses extended pronunciation confusion networks that include both canonical pronunciations and their mispronounced variants [2, 3, 4]; *Confidence-based* approach measures the similarity between native speaker's canonical pronunciation and its corresponding realization by L2 learners [5, 6, 7]. ASR utilizes hidden Markov models (HMMs) to capture temporal information of phones, however, HMMs are still not powerful enough to discriminate sounds that are spectrally similar and differ mainly in duration [8]. For example, HMMs are not quite suitable to distinguish fricatives from plosive release segments since the difference of these two sounds is subtle in the amplitude envelope [9].

Therefore, another line of this research, as illustrated in this paper, will cast pronunciation error identification as a binary classification task that improves discrimination power by detecting distinctive feature errors known to occur with high frequency. Acoustic landmark theory [10] by exploiting quantal nonlinear articulatory-acoustic relationships provides a basis of selecting distinctive features that are suitable for speech recognition [11]. We will leverage this theory further for the task of identifying pronunciation errors.

2. Related Works

The factors that cause high-frequency errors differ for L2 learners from different native language backgrounds. For example, the single biggest pronunciation problem for Spanish-speaking learners of English is that Spanish does not have a distinction between short and long vowels [12], while Japanese-speaking learners can mitigate the affects of vowel duration [13].

Acoustic cues that distinguish error minimal pairs include standard ASR features, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP), but also include many specialized cues that have been found to be useful for particular binary contrasts. Voice onset time (VOT) features are proposed to detect Mandarin Chinese (L2) phonetic errors of aspirated consonants (/p/, /t/, /k/) pronounced by Japanese native speakers [14]. Rate of Rise (ROR) values that are calculated by determining the derivative of the logarithm of root-mean-square energy succeed to discriminate the voiceless velar plosive /k/ from the voiceless velar fricative /x/ in Dutch, since the release of the burst of the plosive causes an abrupt rise in amplitude [8]. As for the mispronounced Dutch vowels, formants (F1-F3) have been used with ASR-based confidence measures are exploited further [15]. Goodness of speech prosody (GOSP) has also been defined [16] in terms of several features including F0, duration, parameters of Fujisaki model, rPVI, and nPVI.

Among these feature representations of mispronounced sounds, statistical models are also explored with the purpose of selecting distinctive features. Stouten et al [17] applied artificial neural network models to extract distinctive features from MFCC features in the context of learning English. Lee et al [18] tried to identify error sounds by leveraging features that

doi: 10.21437/SpeechProsody.2016-51

are learned from neural networks in an unsupervised manner. Hacker et al [19] achieved promising performance based on top-15 distinctive features using the AdaBoost algorithm for the task of detecting English errors made by German children.

Recent application of landmark-based distinctive features in ASR motivated researchers to further explore their utility in pronunciation error detection problems. Quantal nonlinearities in articulatory-acoustic relations provide a theoretical basis for selecting distinctive features, complementary to the empirical foundations of most L2 research [10]. Acoustic landmark theory, first described in [20], has been successfully applied in identifying English pronunciation errors produced by Korean speakers [21, 22]. This guidance of selecting distinctive features is probably suitable for a larger pool of languages other than English only. In the context of Chinese learning as a foreign language, Zhang et al [23, 24] developed a Mandarin Chinese distinctive features system based on the knowledge of acoustics and physiology. Wang et al demonstrated the capability of discrimination between several phonemes on that system by comparing the parameters of perceptual linear prediction (PLP), MFCC and linear prediction cepstral coefficients (LPCC) [25]. However, determining the acoustic landmark positions that best represent categorical phonological distinctions remains a difficult problem, since the acquisition of this knowledge requires large scale experiments of human speech perception [25]. The lack of this knowledge hinders the progress of the application on identifying pronunciation errors.

In this paper, we provide two alternative methods for selecting acoustic landmark positions in L2 Chinese. First, we directly mapped well-founded English landmark theory into Mandarin Chinese, since there exists similar phonetic characteristics between these two languages; second, we define Chinese landmark positions and corresponding distinctive features and acoustic cues by analyzing a large scale corpus of pronunciation error pairs.

3. Description of Data

This large scale corpus of Chinese (L2) speech, referred to as BLCU inter-Chinese speech corpus [26], has collected data from more than 100 speakers. Each speaker read a sentence from 301 daily use sentences. This corpus consists of 64,190 phonemes and 4,631 utterances. The continuous speech of 17 Japanese native speakers (8 males and 9 females) has been phonetically annotated at segment level. The annotators are 6 post-graduate students majoring in phonetics, divided into two groups. The speech data were annotated twice independently by the two groups, with each annotator labeling a continuous 200 utterances on a rotating basis. The speech data were manually transcribed and were automatically aligned into phonetic segments of “initials” and “finals” with human transcriptions using HTK [27]. The absolute agreement (in percentage of matching values) between annotators ranges from 77.0% to 84.3% with the average agreement 80.7%. The correlation coefficients are computed for the phoneme based mispronunciation rates for the two groups with the average correlation ratio 0.78. The 65 kinds of pronunciation error tendencies (PETs) based on articulation-placement and articulation-manner are annotated to represent general erroneous articulation tendencies, including raising, lowering, advancing, backing, lengthening, shortening, centralizing, rounding, spreading, labio-dentalizing, laminalizing, devoicing, voicing, insertion, deletion, stopping, fricativizing, nasalizing, and retroflexing [26].

In this study, we only considered consistent utterances where two annotators are in agreement, and select Top-16 frequent PETs covering 61.28% of total pronunciation errors to perform experiments on identifying pronunciation errors, while ignoring other very rare PETs. These 16 PETs were divided into four categories: spreading, backing, shortening and laminalizing. The final corpus consists of 7,837 phones (error: 1,524, correct: 6,313). The error rate across 16 phones ranges from 6% to 44.1%.

4. Methodology

In the context of pronunciation error identification, *confidence-based* approaches still maintain better performance than *rule-based* methods due to the “coverage of confusion networks” trade-offs [28]. The goodness of pronunciation (GOP) [6] algorithm is probably the most widely used measure in this scope. Our baseline for testing landmark-based pronunciation error detections is GOP using deep neural network triphone acoustic models trained on our large scale corpus [29].

4.1. Goodness of Pronunciation (GOP)

The aim of GOP is to provide a confidence score for each phone in a speech utterance. Given the orthographic transcription and acoustic models that determine the likelihood $P(O^q|q)$ where O^q denotes the acoustic segment aligned with phone $q \in Q$, the GOP score can be calculated by normalizing the log likelihood ratio of phone p compared to its strongest competitor over the number of frames $NF(p)$ in phone p,

$$GOP(p) = \left| \log \left(\frac{P(O^p|p)}{\max_{q \in Q} P(O^q|q)} \right) \right| / NF(p) \qquad (1)$$

We applied maximum mutual information (MMI) estimation to adapt acoustic models using Japanese native speaker’s speech. The numerator and denominator in Eq. (1) are computed by forced alignment with orthographic transcription and an unconstrained phone loop, respectively.

4.2. Acoustic Landmark Theory

Stevens proposed [30, 10] four different candidate landmark locations for English, including vowel peak landmark, oral closure landmark, glide valley landmark in glide-like consonants, and oral release landmark. These four landmark categories were proposed by Stevens to be language-universal, but our studies of Mandarin Chinese suggest other signal events that may have a better claim to be both perceptually salient and phonologically distinctive.

In the task of pronunciation error identification, we could explore error pairs from the development corpus in order to define the acoustic landmark positions and distinctive features. We conduct speech perception experiments in collaboration with experts at BLCU Department of Linguistics, and discovered the distinctive Chinese landmark positions for the phones with Top-16 highest error frequencies in the corpus. As an alternative to these perception-based Chinese landmark candidates, we find correspondences of articulatory-manner and articulatory-place between English and Mandarin Chinese after applying Stevens theory. Table 1 lists the landmark positions signaling 16 Chinese phonemes according to these two different methods of landmark definition.

Table 1: Acoustic landmark positions obtained by Chinese phonetics and English phonetics theory. Phone symbols are IPA (*pinyin* in parens). The fraction number denotes the relative time stamp in the duration.

Phone	Chinese Landmark	English Landmark
ʂ (sh)	following vowel	fricative: start, end
ɖʐ (zh)	coda of consonant	affricate: start, end
ʈʂ (ch)	onset of consonant	affricate: start, end
ɕ (x)	following vowel	fricative: start, end
dʑ (j)	following vowel	affricate: start, end
an (an)	onset (14/30) of vowel	nasal: start,end
y (v)	onset of vowel	vowel: middle
aŋ (ang)	onset (14/30) of vowel	nasal: start, end
iŋ (ing)	onset (17/30) of vowel	nasal: start, end
u (u)	onset of vowel	vowel middle
f (f)	onset of consonant	fricative: start, end
əŋ (eng)	onset of vowel	vowel: start, end
tɕ (q)	onset, nucleus, coda	affricate: start, end
k (k)	following vowel	stop: start
ɻ (r)	whole consonant	fricative: start, end
uɔ (uo)	onset, nucleus, coda	glide: middle

5. Experiments and Results

We compared acoustic landmark features (see Table 1) with GOP-based features using 10-fold stratified cross validation. Evaluations were then performed using a held-out test set.

5.1. Setup

In all experiments, MFCC appended by its acceleration, delta coefficients, and C0 coefficients are extracted. Cepstral mean normalization is applied to compensate long-term spectral effects[1]. Formants[2](F1 - F5) are computed from the signal up to 5500Hz since all test-takers are female. A Hamming window of 25ms was used to chunk short-term stationary signals as frames, and the default frame rate is 10ms. However, many Chinese phones have short durations, and many segments therefore contain less than four frames (e.g. /u/ and /i/ often have one frame). To address the issues of insufficient number of frames, MFCC were recomputed using a frame rate adjusted as necessary between 2ms and 10ms. 20% of the whole corpus was held out as a test set that holds the same proportions of class labels.

Taking into account the imbalanced nature of the training set, we applied support vector machine (SVM) with linear kernel, and assigned weights inversely proportional to the class frequencies as suggested in [31].

5.2. Cross Validation

Figure 1 illustrates the micro-average f1 scores for individual phones over six different features. Red bars denote GOP baseline models that hold reasonable performance for most of the phones except /aŋ/(ang) and /k/(k). In comparison to GOP baseline, all acoustic cues at landmarks outperform GOP measure significantly except for the phones /f/(f), /tɕ/(ch), and /ɻ/(r) due to large overlaps of error bars.

From the comparisons between MFCC features at Chinese landmarks (blue) and English landmarks (green), we observed that for the nasal phones /aŋ/(ang), /iŋ/(ing), and /an/(an) with backing errors, English landmarks outperform Chinese. According to the landmarks definition (see Table 1), Chinese landmarks fall on the onset of vowels while English landmark considered the beginning and end of the consonant, which seems to be a better position for discriminating tongue backing errors in the vowel and consonant. For the fricative phones /ʃ/(sh), /ɕ/(x), and /dʑ/(j), Chinese landmarks locating at the following whole vowel perform worse than English consonant-boundary landmarks, despite the perceptually salient vowel difference that co-occurs with the consonant distinction in Chinese.

Formants features (light and cyan) that are literally expected to disambiguate vowels, seem not to contribute for discriminations for all phones except for the phones /iŋ/(ing) and /ɻ/(r). The aspirated stop phone /k/(k) expresses an interesting pattern that the Chinese landmark (the following vowel) achieves a better f1 score than English (only considering the start of the stop release segment). Aggregating all features together as shown in yellow bars made limited improvements particularly in the case that both Chinese and English landmarks compensate with each other, e.g. fricative phone /ʂ/(sh).

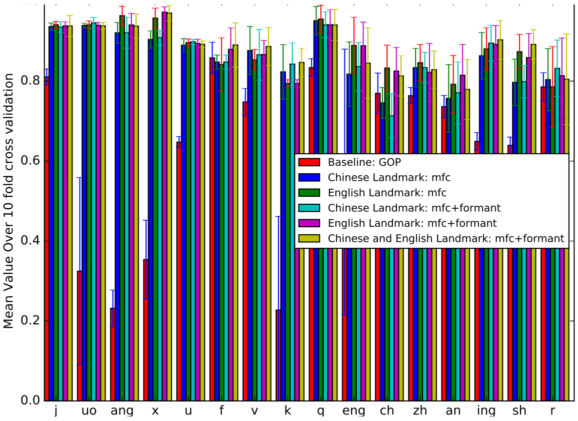


Figure 1: 10-fold cross validation performances of GOP baseline and acoustic features at Chinese and English landmarks. Y-axis shows the micro-average values of f1 scores for individual phones. The sequence of phones is sorted by the error percentage in the training set.

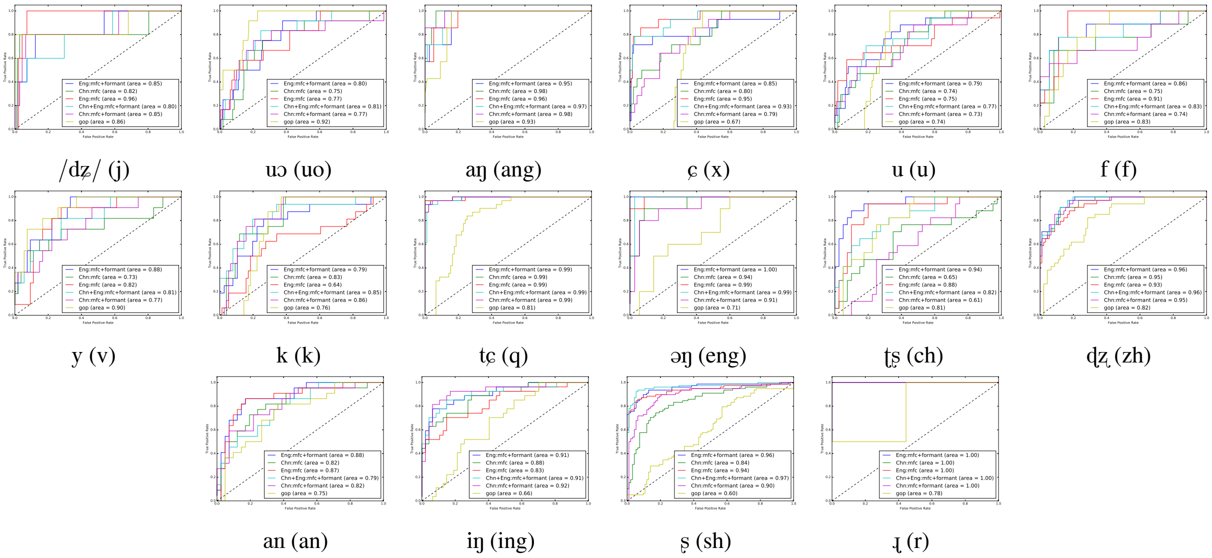
The observation in Figure 1 indicates that acoustic cues beyond MFCC may still be redundant or irrelevant for the classification tasks. For example, formants at the onset of consonant could be irrelevant feature for the affricative phone ʈʂ(ch), and may need to be removed. While for the affricative /dʑ/(j) and glide type /uɔ/(uo), the performance remains unchanged after applying MFCC and formants features at Chinese and English landmarks. Besides of the landmark positions and corresponding acoustic cues, various frame rates are also applied on 16 phones. We empirically choose the best acoustic cues for individual phones based on the best micro-average f1 scores (see Table 2). Seven phones /aŋ/(ang), /y/(v), /tɕ/(q), /əŋ/(eng), /ɖʐ/(zh), /iŋ/(ing), ʂ(sh) can achieve the best performance by using large frame rate (10ms), while the smaller frame rate (4ms) are suitable two phones /uɔ/(uo) and /k/(k).

5.3. Acoustic Cues for Evaluation

Cross validation experiments demonstrate that the performances for individual phones under micro-average f1 score were highly correlated with the combinations of landmark po-

[1] HCopy config parameter: MFCC_0_A_D_Z

[2] http://www.fon.hum.uva.nl/praat/



/dʑ/ (j) uɔ (uo) aŋ (ang) ɕ (x) u (u) f (f)

y (v) k (k) tɕ (q) əŋ (eng) ʈʂ (ch) ɖʐ (zh)

an (an) iŋ (ing) ʂ (sh) ɻ (r)

Figure 2: ROC curves of evaluations on held-out test set for individual phones.

Table 2: Best acoustic cues selected for individual phones.

Phone	FrameRate	Landmark	AcousticCues	F1score
dʑ(j)	6ms	Chn+Eng	mfc+formant	0.949
uɔ(uo)	4ms	Eng	mfc+formant	0.945
aŋ(ang)	10ms	Eng	mfc+formant	0.967
ɕ(x)	6ms	Eng	mfc+formant	0.977
u(u)	8ms	Chn	mfc+formant	0.890
f(f)	8ms	Chn+Eng	mfc+formant	0.902
y(v)	10ms	Chn+Eng	mfc+formant	0.887
k(k)	4ms	Chn	mfc	0.872
tɕ(q)	10ms	Eng	mfc	0.970
əŋ(eng)	10ms	Eng	mfc	0.908
ʈʂ(ch)	8ms	Eng	mfc	0.861
ɖʐ(zh)	10ms	Eng	mfc+formant	0.855
an(an)	6ms	Eng	mfc	0.844
iŋ(ing)	10ms	Eng	mfc+formant	0.919
ʂ(sh)	10ms	Eng+Chn	mfc+formant	0.902
r(r)	8ms	Chn	mfc+formant	0.832

sitions, frame rates, and acoustic cues. We continued to explore the generalization power of these models on our 20% held-out test data. In this study, the best frame rate for each phone was frozen (see Table 2), and six models including GOP baseline were evaluated. In the context of identifying pronunciation errors, L2 learners expect to receive more feedbacks of pinpointing salient errors rather than false alarms. Receiver Operating Characteristic (ROC) metric that formulates the relationship between true positive rate (TPR) and false positive rate (FPR) can mitigate L2 learner's concerns. Error classes was assigned as positive labels. Figure 2 illustrates the curves of comparison results on 16 phones. TPR is on Y-axis, and FPR is on X-axis. This means the top left corner of the plot is the ideal point (TPR=1, FPR=0). Namely, larger area under curve (AUC) indicates better performance. The steepness is also an important sign since the TPR is maximized while keeping FPR minimized.

GOP model (yellow) proved to be a strong baseline for most of the phones, particularly for the phones /uɔ/(uo) and /y/(v) that outperformed all other landmark-based models (AUC>0.9). However, for the phones /ɕ/(x) and /u/(u), the ROC curves of GOP have intercepts with dashed “chance” line, and TPR remains to be zero even when FPR decreases. All other landmark-based models achieved the performance above the “chance” line except for the combined acoustic cues of MFCC and formants at Chinese landmarks on the phone /ʈʂ/(ch).

6. Conclusions

In this paper, we proposed two approaches to select Mandarin Chinese salient phonetic landmarks for the Top-16 frequently mispronounced phonemes by Japanese (L1) learners, and extract features at those landmarks including mel-frequency cepstral coefficients (MFCC) and formants. One is to directly map well-founded English landmark theory into Chinese language since there exists correspondences of articulatory-manner and articulatory-place between English and Mandarin Chinese after applying Stevens theory. Second, we defined distinctive Chinese landmarks for Top-16 frequent pronunciation errors by conducting human speech perception experiments in collaboration with linguists.

In order to make fair comparison, we selected a strong baseline model using goodness of pronunciation (GOP). Experiments including 10-fold cross validation on the training set and evaluation on the held-out test set illustrated that acoustic cues of MFCC and formants at both Chinese landmarks and English landmarks led a better performance significantly. When comparing the performance between these two landmark theory, English landmarks locating at both start and end of durations for most of the 16 phones slightly outperformed Chinese landmarks that was defined by the empirical analysis of error pairs in the large scale corpus. Chinese landmarks might lose some significant information on discriminating pronunciation errors especially for the nasal phones and fricative phones. We expected to get access to even larger corpus that are suitable for us to consolidate our Chinese landmark theory.

7. Acknowledgments

This research was supported by QNRF NPRP 7-766-1-140. Conclusions and findings are those of the authors, and are not endorsed by QNRF. We also express sincere gratitudes to Professor Chilin Shih for the discussions on Chinese phonetics.

8. References

[1] D. Luo, X. Yang, and L. Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus." in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1593–1596.

[2] A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer." in *9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2008, pp. 2787–2790.

[3] W. K. Lo, S. Zhang, and H. M. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system." in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 765–768.

[4] X. Qian, H. M. Meng, and F. K. Soong, "On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (CAPT)." in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 865–868.

[5] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 83–93, 2000.

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

[7] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study." *ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 49–52, 2009.

[8] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[9] L. F. Weigelt, S. J. Sadoff, and J. D. Miller, "Plosive/fricative distinction: the voiceless case," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2729–2737, 1990.

[10] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[11] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan *et al.*, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. I, 2005, pp. 213–216.

[12] P. Iverson and B. G. Evans, "Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 866–877, 2009.

[13] M. Hisagi, K. Nishi, and W. Strange, "Acoustic properties of Japanese and English vowels: Effects of phonetic and prosodic context," *Japanese/Korean Linguistics*, vol. 13, pp. 223–224, 2008.

[14] Y. Gao, R. Duan, J. Zhang, and Y. Xie, "SVM-based mispronunciation detection of Chinese aspirated consonants (/p/, /t/, /k/) by Japanese native speakers," in *Proceedings of 9th Phonetic Conference of China (PCC)*, 2014, pp. 193–196.

[15] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 580–585.

[16] S. Huang, H. Li, S. Wang, J. Liang, and B. Xu, "Automatic reference independent evaluation of prosody quality using multiple knowledge fusions," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 610–613.

[17] F. Stouten and J.-P. Martens, "On the use of phonological features for pronunciation scoring," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, 2006, pp. 329–332.

[18] Y.-B. Wang and L.-S. Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8232–8236.

[19] C. Hacker, T. Cincarek, A. Maier, A. Hebler, and E. Noth, "Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. IV, 2007, pp. 197–200.

[20] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, V. A. Fromkin, Ed. Orlando, Florida: Academic Press, 1985, pp. 243–255.

[21] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based SVM." in *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009, pp. 1903–1906.

[22] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 614–617.

[23] J. Zhang, "Distinctive feature system in Mandarin Chinese," *Chinese Journal of Acoustics*, vol. 30, no. 6, pp. 506–514, 2006.

[24] J. Zhang, "Distinctive feature tree in Mandarin Chinese," *Chinese Journal of Acoustics*, vol. 31, no. 3, pp. 193–198, 2006.

[25] M. Wang and Z. Meng, "Classification of Chinese word-finals based on distinctive feature detection," *Third International Symposium on ElectroAcoustic Technologies (ISEAT)*, vol. 35, no. 9, pp. 38–41, 2011.

[26] W. Cao, D. Wang, J. Zhang, and Z. Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 1922–1925.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book (for HTK version 3.4)*. Cambridge University Press, 2006.

[28] L. Wang, X. Feng, H. M. Meng *et al.*, "Automatic generation and pruning of phonetic mispronunciations to support computer-aided pronunciation training." in *9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2008, pp. 1729–1732.

[29] Y. Gao, Y. Xie, W. Cao, and J. Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 693–696.

[30] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features," in *Second International Conference on Spoken Language Processing (ICSLP)*, vol. 1, Banff, Alberta, 1992, pp. 499–502.

[31] X. Yang, A. Loukina, and K. Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 300–305.