LANDMARK OF MANDARIN NASAL CODAS AND ITS APPLICATION IN PRONUNCIATION ERROR DETECTION

Yanlu Xie[1], Mark Hasegawa-Johnson[2], Leyuan Qu[1], Jinsong Zhang[1]

[1] College of Information Science, Beijing Language and Culture University, Beijing
[2] Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Illinois

ABSTRACT

L2 learners of Mandarin have difficulty learning native-like pronunciation of nasal codas. In order to help them learn native-like pronunciation, we propose to develop targeted classifiers for automatic pronunciation error detection. In this paper, perceptual experiments with modified speech are designed to analyze the exact position of the landmark of a nasal coda. Based on perceptual results from isolated words, we propose that information about nasal coda place of articulation is most dense near a landmark at the center of the nasalized vowel. Landmarks detected in a database of Japanese learners of Mandarin, and classified as correct vs. incorrect using an SVM. The result shows that the detection performance of the SVM+Landmark system is similar to that of a DNN-HMM+MFCC system. When the two systems are combined, an FRR of 4.6% is achieved at DA of 83.9%. This performance is comparable to that of previously developed classifiers for 16 common Mandarin pronunciation errors.

Index Terms— Landmark, nasal coda, pronunciation error detection, computer aided pronunciation training

1. INTRODUCTION

Mandarin nasal codas play an important role in the standard Chinese pronunciation system. There are 16 nasal rhymes which account for 41% of all rhymes in standard Chinese. There are 177 syllables with nasal coda which account for 44% of the syllables in standard Chinese. Nasal coda acquisition is difficult for foreigners. According to statistical reports of 19 projects, 86.3% of Japanese students consider nasal codas the most difficult feature of Chinese pronunciation [1]. Furthermore nasal codas error detection is also difficult in Computer aided pronunciation training (CAPT) systems [2]. It has been argued that weakness of the syllable rhyme is a universal feature, caused by the prosodic dominance of the syllable onset [3].

A prototypical Mandarin nasal rhyme is composed of three segments: the oral vowel segment, the nasalized vowel segment, and the nasal consonant [4]. Classification of the coda consonant by Chinese native speakers depends little on the oral vowel segment [5], but the nasalized vowel segment strongly influences classification of the coda consonant [4][5][6][7]. Repp [7] and Kurowski [8] considered that the nasalized vowel contains as much information as the consonant, or even more [9]. Some researchers have reported that the consonant identity is determined by the pattern of transition from the nasalized vowel into the consonant [5][10][11].

Stevens' theory of landmark-based speech perception [12] proposes two landmarks in a nasal rhyme: the velar opening landmark (between the oral and nasalized vowel segments), and the oral closure landmark (between the nasalized vowel and the nasal consonant), and his theory proposes that the oral closure landmark should dominate processes of phoneme classification and speech synchronization. Strong nasalization of the vowel in Mandarin causes landmark-based classification to be difficult in two ways. First, it is not clear whether classification should be synchronized with respect to the velar opening landmark or the oral closure landmark; second, strong nasalization of the vowel means that the oral closure landmark is not always easy to locate.

In the field of CAPT (computer-assisted pronunciation training), in order to detect the nasal coda errors automatically, MFCC parameters and relative formant parameters are used [2]. These parameters such as second formant, third formant, formant energy, and harmonics could be viewed as cues for nasalization in perceptual experiments [13][14]. However, classification of consonant place of articulation (alveolar /n/ vs. velar /ŋ/) is inaccurate using these parameters. Coda nasal place of articulation is most frequent of the 16 common pronunciation error tendencies (PET) suffered by Japanese learners of Mandarin [15], and would therefore benefit from accurate CAPT.

Oral closure landmarks have been used in automatic speech recognition (ASR) as anchor points for phoneme classification [16], and in CAPT for the detection of pronunciation errors by Korean learners of English [17][18]. In the latter work, errors were detected using 3-frame samples extracted from four different candidate landmark locations: the temporal midpoint of the vowel (estimated location of Stevens' vowel peak landmark [12]), the boundary between the vowel and the consonant (estimated location of the oral closure landmark), the middle of the consonant (estimated location of Stevens' glide valley landmark in glide-like consonants), and the boundary

between the consonant and its following segment (estimated location of the oral release landmark) [18]. These locations were selected without further acoustic analysis, and therefore may not always correspond to the time of the desired articulatory event.

In this paper, landmark-based Mandarin coda nasal CAPT is proposed. The distinctive features and the timing of the landmarks are evaluated with perceptual experiments. The detection results of coda nasal landmarks and the acoustic parameters are combined.

This paper is organized as follows: in section 2, the subsegments of a nasal rhyme are analyzed. Then the landmark of the Mandarin nasal vowel is analyzed. In section 3, the L2 pronunciation errors detection method is presented. The experiments of nasal codas pronunciation error detection in a continuous word database are described in section 4. In section 5, the results are discussed.

2. NASAL CODA LANDMARKS

2.1. Quantal nature and distinctive features of nasals

Articulatory-acoustic relations exhibit quantal nonlinearities that may be exploited, by any given language, as the basis of distinctive features. Stevens and Mou proposed models of nasalization based on the pole and zero positions [5]. The relative amplitudes of spectral pole change categorically at boundaries among the three subsegments of a nasal rhyme, depending on the relative areas of the oral and velopharyngeal port. The velopharyngeal opening is the articulator responsible for the perception of nasality. After measuring F2, a correlate of the degree of tongue fronting, Lin found that F2 at the end-point of the nasalized vowel segment in a nasal rhyme is highly correlated with Mandarin nasal coda place of articulation [4]. Kurowski and Blumstein found [8], however, that the endpoint of a nasalized vowel (the oral closure landmark) is not always easy to locate.

2.2. Landmark perception

In order to find the landmark position and the distinctive features of Mandarin nasal codas, we study the perceptual influences from vowel segments on the judgments of alveolar/velar nasals by native speakers of Chinese. This study follows the protocol proposed in [19].

2.2.1. Materials

A splicing methodology is used to measure the perception result of the three segments of nasal rhymes. Each syllable is divided into four segments: I (initial consonant), V (oral vowel), T (nasalized vowel), and N (nasal consonant). Syllables may be modified by removing (-) segments, or by adding (+) segments from other syllables: t (nasalized vowel) or n (nasal consonant), resulting in three types of modification:

IV+t-N: nasal consonant is cut and nasalized vowel is exchanged, as shown in Fig. 1.

IV-T+N: nasalized vowel is cut, as shown in Fig. 2.

IV-T+n: nasalized vowel is cut and nasal consonant is exchanged, as shown in Fig. 3.

In the three figures, the left two waves are original speech ban1 and bang1. The right two waves are modified speech.

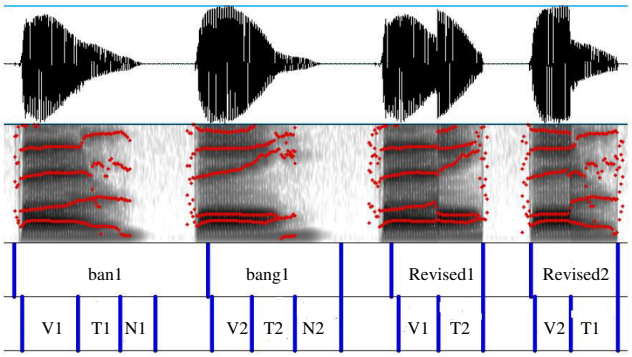


Fig. 1: Modified speech, nasal vowel replaced (IV+t-N)

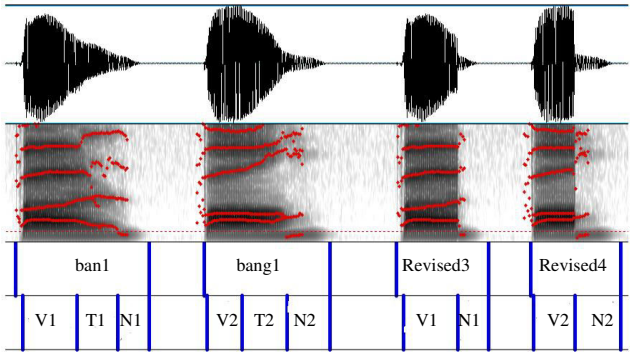


Fig. 2: Modified speech, nasal vowel cut (IV-T+N)

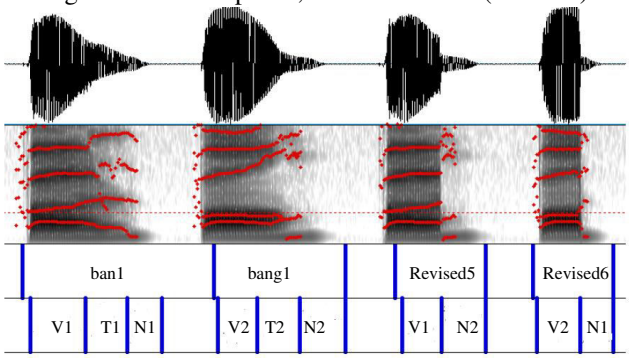


Fig. 3: Modified speech, consonant replaced (IV-T+n)

The original speech data are from a female and a male speaker of the 863 Corpus of Speech Synthesis-1 (CoSS1) [20]. 46 tokens are selected for each of four rhyme types, (an, ang, en, and eng), for a total of 184 (92×2) unmodified and 552 (92×2×3) modified tokens. The three segments of the nasal rhyme are annotated by undergraduate students majoring in experimental phonetics. It is difficult to find the boundaries of the three segments. Lin proposed that the start point of the nasal consonant is the time when the oral articulator closes (in this case, the tongue); at this time, the spectral amplitude and shape change greatly [3]. So the

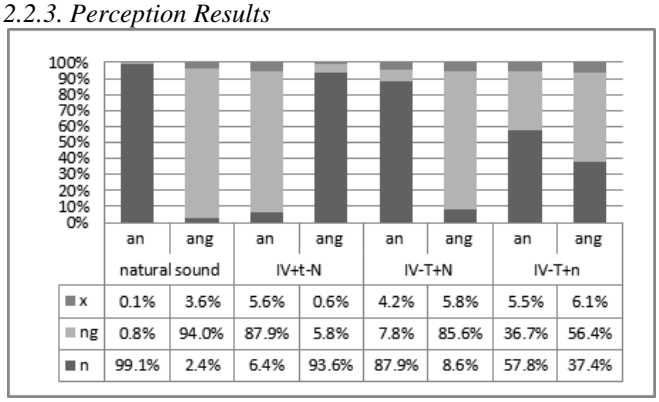
boundaries are considered regarding the formants, spectrum and waveform together.

Syllables of in/ing are not selected in the perception experiments. The pronunciation of /ing/ becomes /iəŋ/ in typical Mandarin speech, with a transitional schwa between oral vowel and nasalized vowel [5], therefore this rhyme pair was omitted from analysis.

2.2.2. Participants and Procedure

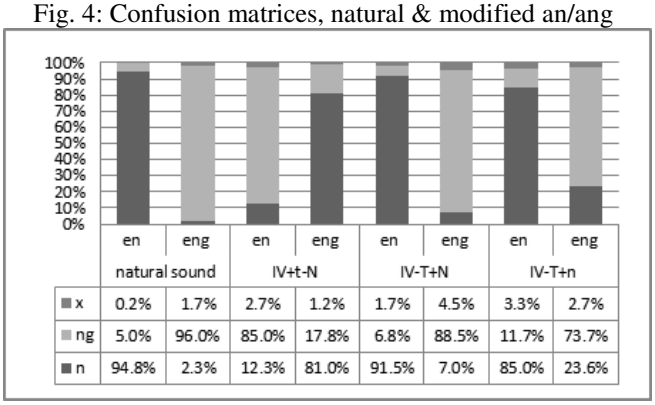
15 native Chinese (7 males and 8 females) participate in the perception experiments. The materials are presented in random order using E-PRIME. The participants are forced to choose from three labels: front nasal n/n/(an, en), back nasal ng/ŋ/(ang, eng), no coda x.

2.2.3. Perception Results



	an	ang	an	ang	an	ang	an	ang
x	0.1%	3.6%	5.6%	0.6%	4.2%	5.8%	5.5%	6.1%
ng	0.8%	94.0%	87.9%	5.8%	7.8%	85.6%	36.7%	56.4%
n	99.1%	2.4%	6.4%	93.6%	87.9%	8.6%	57.8%	37.4%

Fig. 4: Confusion matrices, natural & modified an/ang



	en	eng	en	eng	en	eng	en	eng
x	0.2%	1.7%	2.7%	1.2%	1.7%	4.5%	3.3%	2.7%
ng	5.0%	96.0%	85.0%	17.8%	6.8%	88.5%	11.7%	73.7%
n	94.8%	2.3%	12.3%	81.0%	91.5%	7.0%	85.0%	23.6%

Fig. 5: Confusion matrices, natural & modified en/eng

Fig. 4 and Fig. 5 show percentage of correct and incorrect classification, by human subjects, of an/ang and en/eng respectively. Subjects could indentify natural coda consonants with accuracy above 94%. When nasal consonant is cut and nasalized vowel is exchanged, the recognition results almost totally change (above 81%), implying that the replaced nasalized vowel is sufficient to change perceptual label of the token. When nasalized vowel is cut, classification accuracy is still above 85%, demonstrating that the sequence of oral vowel and nasal consonant is sufficient for accurate classification. When nasalized vowel is cut and nasal consonant is exchanged, the majority of subjects (above 50%) still hear the original consonant, implying that when the oral vowel and nasal consonant are unmatched, subjects do not know how to respond.

The results show that the nasalized vowels play a dominating role in perception of nasal codas an/ang and en/eng. We propose therefore that the acoustic landmark most useful for classification of nasal coda place of articulation is the middle of the nasalized vowel, rather than its beginning or end.

3. NASAL CODAS PRONUNCIATION ERROR DETECTION SYSTEM

The baseline nasal coda pronunciation error detector is a DNN based system. In order to verify the effectiveness of the nasalized vowel, an SVM based landmark system is combined with the baseline system.

3.1. DNN based pronunciation error detection

A set of diacritics were designed for different kinds of common PET [15]. As to the PET, the extended pronunciation network is designed to represent the possible pronunciation variants in the annotation convention. Then a deep neural network (DNN) is used to model the acoustic features of the pronunciation. The DNN is trained in a layer-by-layer manner and the layers are constructed by stacking up multiple Restricted Boltzmann Machines. MFCC features are used in the DNN system. In the nasal codas task, only the nasal diacritics are selected and the nasal results are selected from the DNN score.

3.2 Landmark-SVM pronunciation error detector

A SVM is trained for each nasal coda. Positive and negative examples of each coda are selected from pronunciation variants recorded by a native speaker. SVM inputs include MFCC and formant features. For the continuous speech data, it is impossible to find the boundaries of the three segments of a nasal rhyme, so landmark times are estimated based on proportional segment durations measured using corpus CoSS1. The ratio of landmark time (middle position of the nasalized vowel) divided by total length of the nasal rhyme is 14/30, 12/30, and 17/30 for an, en, and ing respectively. Three frames from the middle position are selected. MFCC+d+dd (39 measurements) and formants (6: F1, F2, F3 and delta formants) from three frames are concatenated to create a 135-dimensional feature vector. Ing is not selected in the perception experiments. But its landmark time can be estimated with the same method of that of an, en.

4. EXPERIMENTS

Nasal rhyme tokens from a large Chinese L2 speech database are used as the experiment data [21]. The database includes 301 frequent utterances of Chinese and 26431 phonemes. It is spoken by 7 female Japanese speakers. It is referred to as BLCU inter-Chinese speech corpus and annotated by 6 undergraduate students.

There are 65 kinds of specific PETs as annotated. Most of them are too rare in the corpus to train the acoustic models. The 16 most common PETs are used to calculate the detection performance. The 16 PET categories can be divided into four kinds: spreading, backing, shortening and laminalizing. This ontology groups nasal coda mispronunciations in the backing category, which is the largest. The three nasal rhymes most commonly mispronounced are an, en and ing accounting for 25.1% of all 16 pronunciation errors.

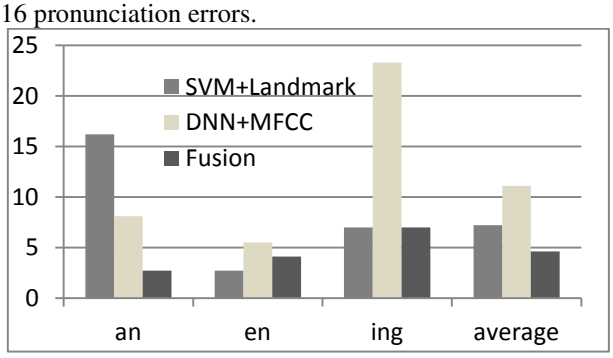


Fig. 6: False rejection rates (FRR %)

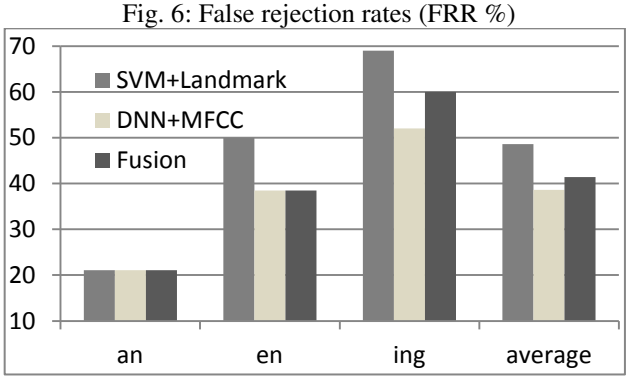


Fig. 7: False acceptance rates (FAR, %)

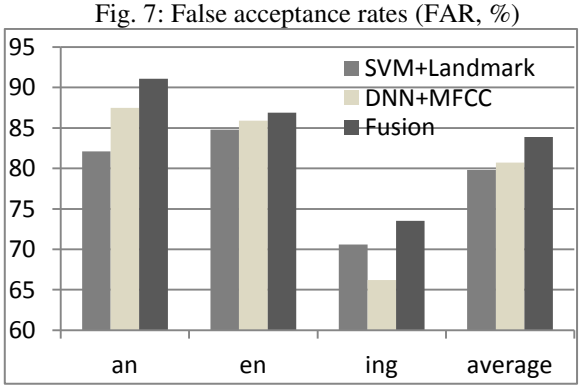


Fig. 8: Diagnostic accuracy (DA, %)

Three kinds of metrics are used to evaluate the detection performance. False Rejection Rate (FRR): The percentage of correctly pronounced phones that are erroneously rejected as mispronounced. False Acceptance Rate (FAR): The percentage of mispronounced phones that are erroneously accepted as correct. Diagnostic Accuracy (DA): The percentage of detected phones that are correctly recognized. While aiming to maximize the DA and minimize both error rates (FAR and FRR), there is an inherent trade-off between the two error rates. Considering the purpose of CAPT, it is critical to avoid discouraging learners by rejecting their correct pronunciations. Therefore DA and FRR are more important in measuring the detection performance than FAR. The detection models and the decision threshold are optimized by aiming at maximizing DA. Due to the fact that phones pronounced correctly are much more that pronounced error in the corpus. FRR is more decisively than FAR in calculating DA.

As described in [15], when detecting the 16 most popular PETs with DNN-HMM+MFCC method, the average FRR = 6.7%, FAR = 35.9%, DA = 87.6%. At the same time, when calculating the detection results of the three nasal rhyme mispronunciations an, en and ing with the same method, the average FRR = 11.1%, FAR = 38.6%, DA = 80.7%. Nasal rhyme errors are diagnosed with lower accuracy than that of any other PET.

Figs. 6, 7, and 8 compare the PET detection performance of the SVM+Landmark and the DNN-HMM+MFCC which was employed in the previous work. The Landmark+SVM has lower FRR than the DNN-HMM+MFCC. On average, the Landmark+SVM system has higher FAR and lower DA than the DNN-HMM+MFCC, but for the particularly difficult "ing" rhyme, the Landmark+SVM improves both FRR and DA, suggesting that the Landmark-based system may be most effective for the most difficult rhyme categories.

Aiming to maximize the DA, the scores of the two systems are combined by voting selection. The combined system outperforms either component system in almost all three metrics (FRR=4.6%, FAR=41.4%, and DA=83.9%), and approaches the average accuracy measures achieved by other PET detector systems in previous work [15].

5. CONCLUSIONS

Nasal coda mispronunciations account for about a quarter of the common mispronunciation errors in L2 Chinese. In order to detect nasal coda mispronunciations automatically, this paper proposed a landmark based method. First, perceptual experiments suggest that the nasalized vowel segment dominates perception; therefore we propose a landmark at the center of the nasalized vowel segment. Detection experiments show that the performance of Landmark+SVM is similar to that of DNN-HMM+MFCC. When the two systems are fused, the performance of nasal coda error detection approaches the average of 16 common PETs.

6. ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of China (61175019), Beijing Higher Education Young Elite Teacher Project (YETP0879), and Research Projects of Beijing Language and Culture University (Special Funds of Basic Research Costs for the National University)(13YBG48).

7. REFERENCES

[1] Shibuya Syuuji. Survey of the difficulties and emphasis for Japanese students learning Chinese [J]. Chinese Language Learning, 2005 (01)

[2] Duan, R., Zhang, J., Cao, W., & Xie, Y. A Preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners[C]. Interspeech. 2014.

[3] S Manuel, S. Shattuck-Hufnagel, K. N. Stevens, R. Carlson and S. Hunnicutt, Studies of Vowel and Consonant Reduction[C],ICSLP 1992, pp. 943-946

[4] Lin, M.C. & Yan, J. Z. Coarticulation in the zero-initial syllable with nasal ending in Standard Chinese. Report of Phonetic Research, Institute of Linguistics,(CASS), Beijing, China, pp. 68-86.1991

[5] Mou X. Nasal codas in Standard Chinese: a study in the framework of the distinctive feature theory [D]. Massachusetts Institute of Technology, 2006.

[6] Recasens D. Place cues for nasal consonants with special reference to Catalan [J]. J. Acoust. Soc. Am, 1983, 73(4): 1346-1353.

[7] Repp B H. Perception of the [m]–[n] distinction in CV syllables [J]. J. Acoust. Soc. Am, 1986, 79(6): 1987-1999.

[8] Kurowski K, Blumstein S E. Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants [J]. J. Acoust. Soc. Am, 1984, 76(2): 383-390.

[9] Recasens D, Marti J. Perception of unreleased final nasal consonants [J]. J. Acoust, 1990, 3: 287-299.

[10] Harrington J. The contribution of the murmur and vowel to the place of articulation distinction in nasal consonants [J]. J. Acoust. Soc. Am, 1994, 96(1): 19-32.

[11] Ohde R N. The development of the perception of cues to the [m]–[n] distinction in CV syllables [J]. J. Acoust. Soc. Am, 1994, 96(2): 675-686.

[12] Stevens K N. Acoustic phonetics[M]. MIT press, 2000.

[13] Lai Y. Acoustic correlates of Mandarin nasal codas and their contribution to perceptual saliency [J]. Concentric: Studies in Linguistics, 2009, 35(2): 143-166.

[14] Chen M Y. Acoustic analysis of simple vowels preceding a nasal in Standard Chinese [J]. Journal of Phonetics, 2000, 28(1): 43-67.

[15] Yingming Gao, Yanlu Xie, Wen Cao, Jinsong Zhang. A Study on Robust Detection of Pronunciation Erroneous Tendency Based on Deep Neural Network[C], Interspeech. 2015.

[16] Hasegawa-Johnson M, Baker J, Borys S, et al. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop[C]. ICASSP., 2005, 1(1415088): 1213.

[17] Yoon S Y, Hasegawa-Johnson M, Sproat R. Automated pronunciation scoring using confidence scoring and landmark-based SVM[C] Interspeech. 2009: 1903-1906.

[18] Yoon S Y, Hasegawa-Johnson M, Sproat R. Landmark-based automated pronunciation error detection[C] Interspeech. 2010: 614-617.

[19] Wang Z, Zhang J. Influences of vowels on perception of nasal codas in Mandarin for Japanese learners and Chinese[C] Chinese Spoken Language Processing (ISCSLP), 2014: 433-433.

[20] CAI Lian-hong, CUI Dan-dan, CAI Rui. TH-CoSS,a Mandarin Speech Corpus for TTS[J] JOURNAL OF CHINESE INFORMATION PROCESSING vol21-2 94-99 2007.3

[21] 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[C], INTERSPEECH, 2010