ANALYSIS ON MISPRONUNCIATIONS IN CAPT BASED ON COMPUTATIONAL SPEECH PERCEPTION

Jia Jia[1,2,3], *Wai-Kim Leung*[4], *Ye Tian*[1,2,3], *Lianhong Cai*[1,2,3] *and Helen M. Meng*[4, 5]

[1]Key Laboratory of Pervasive Computing, Ministry of Education
[2]Tsinghua National Laboratory for Information Science and Technology (TNList)
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
[5]Shun Hing Institute of Advanced Engineering
The Chinese University of Hong Kong, Hong Kong SAR

{jjia, ye-tian10, clh-dcs}@tsinghua.edu.cn, {wkleung, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Computer-aided Pronunciation Training (CAPT) technologies enable the use of automatic speech recognition to detect mispronunciations in second language (L2) learners' speech. In order to further facilitate learning, we aim to be able to develop a principle-based method for generating a gradation of the severity of mispronunciations. This paper presents an approach towards gradation that is motivated by auditory perception. We have developed a computational method for generating a perceptual distance (PD) between two spoken phonemes. This is used to compute the distance between two phonemes of a target (L2) language. The PD is found to correlate well with the mispronunciations detected in CAPT system for Chinese learners of English, i.e. L1 being Chinese (Mandarin) and L2 being US English. These results indicate that auditory confusion indirectly reflects pronunciation confusions in L2 learning. The PD can also be used to help us grade the severity of errors (i.e. mispronunciations that confuse more distant phonemes are more severe) and accordingly prioritize the order of corrective feedback generated for the learners.

Index Terms — second language learning, computer-aided pronunciation training, mispronunciation, computational speech perception

1. INTRODUCTION

The growing number of second language (L2) learners creates a large demand of language learning resources. It is estimated that the English learners in India and China is over 500 million [1] which is greater than the combined population of English speaking countries. This creates a serious of shortage of professional English teachers. A computer-aided pronunciation training (CAPT) system is one of the best approaches to supplement the demands. The traditional recognizer aims for language modeled-constrained lexical training instead of mispronunciation training. In other words, the recognizer still gives out correct words even when the speech contains mispronunciation. We enhance the recognizer with an extended pronunciation lexicon (ERN) [2] to enable pronunciation variation detection and diagnosis. The ERN is generated from phonological rules or a data-driven approach [2][6][7] which includes the common mispronunciations of Chinese speakers. Our group has developed an online CAPT system, Enunciate [3], with an enhanced recognizer for mispronunciation detection and diagnosis, and a synthesizer for corrective feedback generation. The system is now available within The Chinese University of Hong Kong (CUHK) campus and has been used by the hundreds of students and their teachers.

Most of the learners show appreciation of automatic mispronunciation detection technologies. To help learners focus on their pronunciation problems more easily, a gradation of mispronunciations will be helpful. For example, if a learner mispronounces */ih t/* as */ix t/* and */f ae n/* as */f a n/*, the system should show that the substitution error of */ae/ → /a/* is more salient then */ih/ → /ix/*. This gradation can act as a suggestion of priority for the learner to practice their pronunciations. As effective speech communication relies on both the speech production and auditory perception, we suggest that the gradation of mispronunciation should be based on not only the mispronunciation statistics but also a perceptual analysis between two phonemes. There have been considerable research efforts on the computational methods of speech perception for Chinese [4] [5], which allow us to establish a method for analyzing mispronunciation in CAPT by both pronunciation statistics and auditory perception.

In this paper, we propose a method for analyzing mispronunciations in CAPT based on computational speech perception, in order to derive a gradation of the severity of mispronunciations in L2 speech. We begin by presenting a formulation of the problem. Then we take Chinese speakers learning English as an example, giving the statistical results

of mispronunciation from the Enunciate system. Next, we discuss the computational method to generate the "perceptual distance" between English phonemes. Finally, the correlation between mispronunciation statistics and perceptual distances are experimentally investigated, which leads to some suggestions on a priority to for practicing pronunciations targeted at Chinese learners of English.

2. PROBLEM FORMULATION

As effective communication relies on speech production and speech perception, our motivation focuses on how to establish a method for analyzing mispronunciations in L2 speech by both mispronunciation statistics and speech perception.

We use an L2 English corpus with recordings from Chinese learners, which has been phonetically labeled by a trained linguist. Deviations between the labeling and dictionary-based pronunciations form the observed mispronunciations. We define the correct rate of pronunciation as well as the rate of mispronunciation for a given phoneme as follows:

Definition 1: For a phoneme p, the correct pronunciation rate $C(p)$ is the percentage of all correctly pronounced p in all occurrences of p in the L2 training material.

Definition 2: For a phoneme p, the mispronunciation rate $M(p_m, p)$ is the percentage of the mispronunciation $p \rightarrow p_m$ in all occurrences of p in the L2 training materials.

Note that the relationship between C and M is:

$$\sum_{p_m \neq p} M(p_m, p) + C(p) = 1 \qquad (1)$$

In addition, we present the *perceptual distance* (PD) proposed in our previous work [4,5] to compute and evaluate the perceived auditory distance between two phonemes. We then investigate whether there is correlation between the *PD* across different phonemes and the rates of mispronunciation. Should such correlation exist, we aim to utilize the *PD* to derive a gradation of the severity of observed mispronunciations. The gradation should hence be perceptually motivated, and can also be used to generate a prioritization for corrective feedback generation in the context of (computer-aided) pronunciation training applications.

We take Chinese speakers learning English as an example. The CU-CHLOE corpus [6] provides observations on learners' pronunciation. Statistics of detected mispronunciations are shown in next section.

3. MISPRONUNCIATION STATISTICS

We use CU-CHLOE as our corpus which includes the spoken utterances of 100 (50 females and 50 males) Chinese speakers learning English [8]. The training materials of this corpus include:

(i) the Aesop's Fable "The North Wind and the Sun", which has six sentences and cover all the English phonemes;
(ii) a set of 20 phonemic sentences designed by English teachers to cover the common English mispronunciation;
(iii) a set of 10 pairs of confusing words from the TIMIT;
(iv) a set of 50 pairs of minimal pairs from the TIMIT.

Table 1: English vowels with the lowest rates of correct pronunciation *C*, based on CU-CHLOE corpus. Phonemes in boldface exist in American English but not in Chinese.

Phoneme	#Total	# Correct	Correct rate (*C*)	Rate of common mispronunciations (*M*)
/er/	7100	889	17.4%	**/ax/**(40.7%), /ee/ (15.1%)
/aa/	7300	3545	48.6%	/ao/ (38.3%), **/ax/** (5.4%)
/ax/	13900	7498	54.0%	/_/ (12.6%), **/ix/** (11.5%), **/ux/** (7.5%)
/ih/	8700	5054	58.1%	**/ix/** (27.9%), /iy/ (6.6%)
/uh/	1200	829	69.1%	/uw/ (23.3%), **/ux/** (6.8%)
/eh/	4400	3054	69.4%	**/ae/** (17.1%), /ih/ (3.4%)
/ae/	6600	4721	71.6%	**/aa/** (13.7%), **/ax/** (8.2%), /eh/ (4.2%)

Table 2: English consonants with the lowest rates of correct pronunciation *C*, based on CU-CHLOE corpus. Phonemes in boldface exist in American English but not in Chinese.

Phoneme	#Total	# Correct	Correct rate (*C*)	Rate of common mispronunciations (*M*)
/r/	10800	5516	51.1%	/_/ (35.8%), /w/ (5.7%)
/z/	4400	2551	58.0%	/s/ (38.3%), /_/ (2.9%)
/th/	1200	702	58.5%	/f/ (37.7%), /_/ (1.6%)
/jh/	500	299	59.8%	/_/ (13.0%), **/ch/** (12.0%), **/zh/** (4.4%), **/sh/** (3.6%)
/v/	3100	1899	61.3%	/f/ (32.7%), /w/ (3.9%)
/d/	8200	5452	66.5%	/_/ (15.3%), **/t/** (10.7%)
/dh/	6100	4189	68.7%	/d/ (23.6%), /_/ (2.8%)

Table 3: English vowels with the highest rates of correct pronunciation *C*, based on CU-CHLOE corpus. Phonemes in boldface exist in American English but not in Chinese.

Phoneme	#Total	# Correct	Correct rate (*C*)
/oy/	1200	1177	93.1%
/ao/	4300	3935	91.5%
/ay/	4800	4145	86.4%
/aw/	900	758	84.2%
/uw/	3300	2717	82.3%
/ow/	3100	2529	81.6%
/ah/	3600	2772	77.0%

Table 4: English consonants with the highest rates of correct pronunciation *C*, based on CU-CHLOE corpus. Phonemes in boldface exist in American English but not in Chinese.

Phoneme	#Total	# Correct	Correct rate (*C*)
/b/	4200	4166	99.2%
/f/	4000	3964	99.1%
/hh/	5800	5744	99.1%
/g/	1900	1876	98.8%
/w/	4400	4300	97.7%
/sh/	2100	2031	96.8%
/s/	10600	10081	95.1%

Tables 1 and 2 show the lowest rates of correct pronunciation for vowels and consonants in English. For the

phonemes exist only in English and not in Chinese, they have a higher chance of being mispronounced as other similar phonemes by Chinese speaker. We take /er/ as an example. As /er/ does not exist in Chinese, only 17.4% of its occurrences are pronounced correctly but 40.7% are mispronounced as /ax/. This phenomenon is more obvious in consonants. For the seven English consonants with the lowest rates of correct pronunciation, five of them do not exist in Chinese and these phonemes are deleted or substituted by other phonemes which exist in Chinese. On the contrary, if the phonemes exist in Chinese and English, Chinese speakers tend to pronounce them correctly. Tables 3 and 4 show the highest rates of correct pronunciation the English vowels and consonants respectively.

4. AUDITORY PERCEPTUAL DISTANCE

The computational methods of obtaining the vowel perceptual distance and consonant perceptual distance are described in this section.

4.1 Vowel Perceptual Distance

In our previous work, we have proposed a method [4] based on LPC (Linear Predictive Coding) spectral coefficient to measure the differences of vowel perception. By agglomerative hierarchical clustering to LPC features, the vowels are classified by groups. The distances between vowels in the same group are smaller than those in different groups. Compared to vowel perception testing results, we find that vowels in the same group could be confused more easily. This means that the closer the distance, the smaller the difference of an auditory perception. The distance measurement method works well in vowel perception discrimination.

Table 5: the rates of pronunciation and perceptual distances for /er/ and /ey/.

Phoneme	Pronounced as	Rate of pronunciation	Perceptual distance (*PD*)
/er/	/er/	17.4%	0
/er/	/ax/	40.7%	0.5145
/er/	/axr/	9.6%	0.5989
/er/	/eh/	6.6%	0.6783
/ey/	/ey/	75.6%	0
/ey/	/eh/	8.5%	0.4666
/ey/	/ih/	5.1%	0.5384

In the same way, we extract the LPC features to calculate the vowel perceptual distance for vowels in the English training corpus of Enunciate system. The Euclidean distance is selected to measure the perception distance between different vowels. Take vowels /er/ and /ey/ as examples – The results are shown in Table 5. The perception distance is normalized to the interval [0, 1]. As Table 5 shows, the smaller the perceptual distance, the higher the mispronunciation rate.

4.2 Consonant Perceptual Distance

We also studied consonant perception [5]. By analyzing and clustering on consonant features, such as duration, short time energy, zero-crossing rate, MFCC and Bark Rate, the consonants are classified in groups. After removing the redundant features, the dimension of feature vector is compressed to 17 [5]. Compared with the confusion matrix of perception test, we find that consonants in the same group could be confused more easily. That means the closer the distance, the smaller the difference of consonant perception. The distance measurement method works well in consonant perception discrimination.

So for the consonants in English training corpus of Enunciate system, 17 features were extracted. The specific steps are:

- Extract zero-crossing rate according to the formula:

$$Z_n = \frac{1}{2}\sum_{i=2}^{N}\left|\mathrm{sgn}(x_n(i)) - \mathrm{sgn}(x_n(i-1))\right| \quad (2)$$

  where Z_n is zero-crossing rate, N is the number of sample points of the current analysis frame, $\mathrm{sgn}(x_n(i))$ is the sign of the ith sample point in the nth frame;
- Extract the Mel Frequency Cepstral Coefficient (MFCC), and selecting 6 MFCCs as [5] shows;
- Extract Bark Rate
  (i) Calculate the FFT power spectrum;
  (ii) Calculate the integral of the FFT power spectrum for each of 21 Bark bands, marked as x_1, $x_{2,\ldots}x_{21}$;
  (iii) Calculate the Bark Rate y_i :

$$y_i = x_i \Big/ \sum_{j=1}^{21} x_j \quad (3)$$

  (iv) Select 10 Bark Rate features as shows;

Thus, we get all the 17-dimensional feature vectors for each consonant.

- Calculate the distances between consonant feature vectors, and the Euclidean distance is selected to measure the perception distance between consonants.

Table 6: the rates of pronunciation and perceptual distances for /r/.

Phoneme	Pronounced as	Rate of pronunciation	Perceptual distance (*PD*)
/r/	/r/	51.1%	0
/r/	_	35.8%	0.4047
/r/	/ax/	6.1%	0.4686
/r/	/w/	5.7%	0.5233

Take the consonant /r/ as an example – The result is shown in Table 6. The perception distance is normalized to the interval [0,1]. As Table 6 shows, the smaller the perceptual distance, the higher the mispronunciation rate.

5. EXPERIMENTAL RESULTS

Next, we experimentally investigate the correlation between mispronunciation statistical results and the auditory perceptual distances.

5.1 Analysis of phonemes with high correct pronunciation rates

First, considering the seven vowels and consonants with the highest pronunciation correct rates, we calculate the average

perceptual distance $APD(x_i)$ between the vowel/consonant and all the other vowels/consonants. The formula is described as follows:

$$APD(x_i) = \frac{1}{N-1}\sum_{j \neq i} dist(x_i, x_j) \quad (4)$$

where x_i represents for a vowel (or a consonant), $APD(x_i)$ is the average perceptual distance between x_i and the other vowels (or consonants), N is the total number of vowels (or consonants), $dist(x_i, x_j)$ represents the perceptual distance between x_i and x_j. The results are shown in Tables 7 and 8.

Table 7: The average perceptual distances of English vowels with highest rates of correct pronunciation.

Phoneme	Correct rate (*C*)	Average perceptual distance (*APD*)
/oy/	93.1%	0.9035
/ao/	91.5%	0.9146
/ay/	86.4%	0.8803
/aw/	84.2%	0.8526
/uw/	82.3%	0.8631
/ow/	81.6%	0.7406
/ah/	77.0%	0.7367

Table 8: The average perceptual distances of English consonants with highest rates of correct pronunciation.

Phoneme	Correct rate (*C*)	Average perceptual distance (*APD*)
/b/	99.2%	0.9140
/f/	99.1%	0.9246
/hh/	99.1%	0.8923
/g/	98.8%	0.8526
/w/	97.7%	0.8631
/sh/	96.8%	0.8406
/s/	95.1%	0.8467

As shown in Tables 7 and 8, the average perceptual distances of phonemes with highest rate of correct pronunciation are quite high (compared with the results in Table 9 and 10). In general, the *APD* and correct rate *C* are in direct proportion. The higher the *APD*, the more difficult the phoneme is to be confused as another phoneme, which leads to a higher correct pronunciation rate *C*.

5.2 Analysis of phonemes with low correct pronunciation rates

Table 9: The *APD* of English vowels with the lowest rates of correct pronunciation.

Phoneme	Correct rate (*C*)	Rate of common mispronunciation (*M*)	Average Perceptual Distance (*APD*)
/er/	17.4%	/ax/(40.7%), /ee/(15.1%)	0.5137
/aa/	48.6%	/ao/ (38.3%), /ax/ (5.4%)	0.5567
/ax/	54.0%	/_/ (12.6%), /ix/ (11.5%), /ux/ (7.5%)	0.4829
/ih/	58.1%	/ix/ (27.9%), /iy/ (6.6%)	0.4911
/uh/	69.1%	/uw/ (23.3%), /ux/ (6.8%)	0.6667
/eh/	69.4%	/ae/ (17.1%), /ih/ (3.4%)	0.7450
/ae/	71.6%	/aa/ (13.7%), /ax/(8.2%), /eh/ (4.2%)	0.6849

Table 10: The *APD* of English consonants with the lowest rates of correct pronunciation.

Phoneme	Correct rate (*C*)	Rate of common mispronunciation (*M*)	Average Perceptual Distance (*APD*)
/r/	51.1%	/_/ (35.8%), /w/ (5.7%)	0.4640
/z/	58.0%	/s/ (38.3%), /_/ (2.9%)	0.5637
/th/	58.5%	/f/ (37.7%), /_/ (1.6%)	0.5124
/jh/	59.8%	/_/ (13.0%), /ch/(12.0%), /zh/ (4.4%), /sh/ (3.6%)	0.5145
/v/	61.3%	/f/ (32.7%), /w/ (3.9%)	0.4130
/d/	66.5%	/_/ (15.3%), /t/ (10.7%)	0.6304
/dh/	68.7%	/d/ (23.6%), /_/ (2.8%)	0.5560

For the phonemes with low correct pronunciation rates, we also calculate their average perceptual distance. Take /er/ as an example. Since /er/ is easily confused with /ax/ or /ee/, we calculate the average perceptual distance *APD* (/er/) as:

$$APD(/er/) = 1/2[dist(/er/, /ax/) + dist(/er/, /ee/)] \quad (5)$$

As shown in Tables 9 and 10, the *APD* and correct rate *C* are positively correlated. The lower the *APD*, the easier the phoneme is to be confused with other phonemes, which leads to a lower pronunciation correct rate *C*. These results are consistent with the analysis in Section 5.1.

In summary, key observations from the above analysis include: a) The mispronunciation rate *M* and auditory perception distance have negative correlation, which indicates that the auditory confusion will indirectly reflect the spoken confusion; b) For Chinese speakers to learn English, we suggest that learners need to pay more attention to the following phonemes:

/aa/ => /ao/	/ax/ => /ix/
/d/ => /t/	/dh/ => /d/
/eh/ => /ae/	/er/ => /ee/, /axr/,/eh/
/ih/ => /ix/	/th/ => /f/
/v/ => /f/	/z/ => /s/

Where ("A=>B,C,…" means A are most easily to be mispronounced as B or C or …)

6. CONCLUSIONS

This paper studies the problem of analyzing mispronunciation in CAPT. We take Chinese speakers learning English as an example, and propose a novel method for analyzing mispronunciation based on computational speech perception. We discuss the computational method of obtaining the auditory perceptual distance between phonemes. The correlation between mispronunciation statistics and perceptual distances are experimentally investigated. This study indicates that the perceptual distance *PD* and correct pronunciation rate *C* are positively correlated. This also leads to some suggestions on prioritization of corrective feedback generation for CAPT. Future works will focus on how to deal with insertion or deletion in mispronunciation.

7. ACKNOWLEDGEMENT

This work is partially supported by the National Basic Research Program (973 Program) of China (2013CB329304), the research funds from the National Natural Science Foundation of China (61003094, 60931160443), and also funded by the Innovation Fund of Tsinghua-Tencent Joint Laboratory. This project is also partially sponsored by a grant from the CUHK Shun Hing Institute of Advanced Engineering.

8.REFERENCES

[1] B. B. Kachru, *Asian Englishes: Beyond the Canon*, Hong Kong: Hong Kong University Press, 2005.

[2] A. M. Harrison, W. Y. Lau, H. Meng and L. Wang, "Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-sensitive Phonological Rules based on Language Transfer," in *the Proceedings of Interspeech*, September 2008.

[3] K. W. Yuen, W. K. Leung, P. F. Liu, K. H. Wong, X. Qian, W. K. Lo and H. Meng, "Enunciate: An Internet-Accessible Computer-Aided Pronunciation Training System and Related User Evaluations," *in the Proceedings of Oriental COCOSDA - International Conference on Speech Databases and Assessment*, October, 2011.

[4] G. Huang, J. Jia and L. Cai, "A Study on Perception Measurement of Mandarin Vowels Based on LPC Spectrum Features," in *the Proceedings of Phonetic Conference of China May*, 2010.

[5] J. Jia, Y. Wang, Y. Zhang, Y. Tian and L. Cai, "A Discussion on Perception Definition Computing Method of Mandarin Consonants," in *the Proceedings of Phonetic Conference of China*, May, 2012.

[6] H. Meng, Y. Lo, L. Wang and W. Y. Lau, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons," in *the Proceedings of Automatic Speech recognition and Understanding Workshop*, December 2007.

[7] W. K. Lo, A. M. Harrison, H. Meng and L. Wang, "Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-dependent Pronunciation Scoring," in *the Proceedings of the 6th International Symposium on Chinese Spoken Language Processing*, Kuming, China, 2008.

[8] H. Meng, A. Harrison and L. Wang, "Developing a Computer-Aided Pronunciation System for Chinese-Speaking Learners of English," *Bulletin of Advanced Technology Research*, Vol. 3, No. 2, February 2009.