Jia J, Leung WK, Wu YH *et al.* Grading the severity of mispronunciations in CAPT based on statistical analysis and computational speech perception. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(5): 751–761 Sept. 2014. DOI 10.1007/s11390-014-1465-2

Grading the Severity of Mispronunciations in CAPT Based on Statistical Analysis and Computational Speech Perception

Jia Jia[1,2,3] (贾　珈), *Member*, *CCF*, *ACM*, *IEEE*, Wai-Kim Leung[1,4,5] (梁伟俭), Yu-Hao Wu[1,2,3] (吴育昊)
Xiu-Long Zhang[1,2,3] (张秀龙), Hao Wang[4,5] (王　昊), Lian-Hong Cai[1,2,3] (蔡莲红)
and Helen M. Meng[4,5] (蒙美玲), *Fellow*, *IEEE*

[1] *Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

[2] *Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China*

[3] *Key Laboratory of Pervasive Computing, Ministry of Education, Beijing 100084, China*

[4] *Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management The Chinese University of Hong Kong, Shatin, Hong Kong, China*

[5] *Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen 518055, China*

E-mail: jjia@tsinghua.edu.cn; wkleung@se.cuhk.edu.hk; haohaowu2000@163.com; zhangxl13@mails.tsinghua.edu.cn
hwang@se.cuhk.edu.hk; clh-dcs@tsinghua.edu.cn; hmmeng@se.cuhk.edu.hk

Received March 16, 2014; revised July 14, 2014.

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 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 auditory confusion of native language (L1). 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 Cantonese) and L2 being US English. The results show that auditory confusion is indicative of pronunciation confusions in L2 learning. 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.

Keywords 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 number of English learners in China and India is over 500 million[1], which is greater than the combined population of English speaking countries. This creates a serious shortage of professional English teachers. A computer-aided pronunciation training (CAPT) system is one of the best approaches to fulfill the demands. The traditional recognizer aims for language model 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[2-4] which includes the common mispronunciations of Chinese speakers. Our group has developed an online CAPT system, Enunciate[5], with an enhanced recognizer for mispronunciation detection and diagnosis, and a synthesizer for corrective feedback generation. The

Regular Paper

This work is supported by the National Basic Research 973 Program of China under Grant No. 2013CB329304, the National Natural Science Foundation of China under Grant No. 61370023, and the Major Project of the National Social Science Foundation of China under Grant No. 13&ZD189. This work is also partially supported by the General Research Fund of the Hong Kong SAR Government under Project No. 415511 and the CUHK Teaching Development Grant.

system is now available across the campus of The Chinese University of Hong Kong (CUHK) and has been used by hundreds of students and their teachers.

Most of the learners show appreciation for automatic mispronunciation detection technologies. To help learners focus on their pronunciation problems more easily, a gradation of mispronunciations will be helpful[6-7]. 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 than /*ih*/ → /*ix*/. This gradation can act as a suggestion of the priority for the learner to practice specific pronunciations. Wang *et al.*[8] defined several categories of mispronunciation gradations according to different levels of severity. Based on the crowdsourced ratings on mispronunciations of L2 English speech from a large number of listeners, they captured the phonetic context of mispronunciations by phonological rules, which was then augmented with statistical scoring to quantitatively model gradations of word-level mispronunciations. As effective speech communication relies on both 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 has been considerable research on the computational methods of speech perception for Chinese[9-10], which allows 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 generating a gradation of the severity of mispronunciations in L2 speech based on analyzing both mispronunciations in CAPT and computational speech perception. We begin by presenting a formulation of the problem. Then we take Chinese (L1) speakers learning English (L2) 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 two phonemes. Finally, the correlation between mispronunciation statistics and perceptual distances is experimentally investigated, which leads to some principled suggestions on a prioritization for practicing pronunciations targeted at Chinese learners of English.

2 Problem Formulation

We conduct our corpus-based investigation on the L2 English speech of Chinese learners. The corpus has been phonetically labeled by a trained linguist. Deviations between the labeling and the 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:

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

In addition, we present the proposed perceptual distance (PD) to compute and evaluate the perceived auditory distance between two phonemes[4-5]. We then investigate whether there is a correlation between the PD across different phonemes and the rates of various phonetic mispronunciations. 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.

3 Mispronunciation Statistics

3.1 Corpus

The CU-CHLOE (Chinese Learners of English) corpus[3] includes the spoken utterances of 100 Chinese (L1) speakers learning English and has been phonetically labeled by a trained linguist. The deviation can be observed by comparing the labeling and dictionary-based pronunciations①. The corpus includes:

1) the Aesop's Fable "The North Wind and the Sun", which has six sentences and covers all the English phonemes,

2) a set of 20 phonemic sentences designed by English teachers to cover the common English mispronunciation,

3) a set of 10 pairs of confusing words from the TIMIT②,

4) a set of 50 pairs of minimal pairs from the TIMIT.

3.2 Mispronunciation Statistics

Tables 1 and 2 list the English (L2) vowels and consonants with the highest rates of correct pronunciation

①The CMU pronouncing dictionary, www.speech.cs.cmu.edu/cgi-bin/cmudict, July 2014.

②http://catalog.ldc.upenn.edu/LDC93s1, Aug. 2014.

in the corpus. For the phonemes that exist in both Chinese and English, Chinese speakers tend to pronounce them correctly. On the contrary, for the phonemes that exist only in English but not in Chinese, the Chinese learners have a higher probability of mispronouncing them as other similar phonemes. Take /er/ as an example — as it does not exist in Chinese, only 17.4% of its occurrences are pronounced correctly but 40.7% are mispronounced as /ax/. Similarly, five out of seven English consonants with the highest mispronunciation rates exist in English only. These consonants are usually deleted or substituted by other consonants. Tables 3 and 4 list the lowest correct pronunciation rates of the English vowels and consonants respectively.

Figs. 1 and 2 show the complete mispronunciation matrix of Chinese speakers speaking English. Phonemes highlighted in yellow are the phonemes present in English but absent from Chinese according to the manner and place of articulation[11]. For both English vowels and consonants, native Chinese speakers tend to mispronounce the English phonemes as other similar phonemes.

For the English phoneme /l/, it has two sounds called as dark [ɬ] (/el/) and light [ɬ](/l/). A dark [ɬ] comes at the end of a syllable such as bottle (/b aa t ah **l**/). A light [ɬ] comes at the beginning of a syllable such as "late" (/**l** ey t/). In the dictionary, both dark [ɬ] and light [ɬ] share the same alphabet /l/. This causes some occurrence of /l/ to be wrongly marked as mispronunciation when pronouncing as /el/. That is the reason for the low correct pronunciation rate of the phoneme /l/ even though it is present in both Chinese and English.

Table 1. English Vowels with the Highest Rates of Correct Pronunciations C Based on the CU-CHLOE Corpus

Phoneme	Total Number of Pronounciations	Number of Correct Pronounciations	Correct Rate (C) (%)
/oy/	1 200	1 177	93
/ao/	4 299	3 935	92
/ay/	4 799	4 145	86
/aw/	899	758	84
/uw/	3 290	2 717	83
/ow/	3 099	2 529	82
/ah/	3 589	2 772	77

Note: Phonemes in boldface exist in American English but not in Chinese.

Table 2. English Consonants with the Highest Rates of Correct Pronunciations C Based on the CU-CHLOE Corpus

Phoneme	Total Number of Pronounciations	Number of Correct Pronounciations	Correct Rate (C) (%)
/b/	4 199	4 166	99
/f/	3 999	3 964	99
/hh/	5 794	5 744	99
/g/	1 899	1 876	99
/w/	4 357	4 300	99
/sh/	2 099	2 031	97
/s/	10 599	10 081	95

Note: Phonemes in boldface exist in American English but not in Chinese.

Table 3. English Vowels with the Lowest Rates of Correct Pronunciations C Based on the CU-CHLOE Corpus

Phoneme	Total Number of Pronounciations	Number of Correct Pronounciations	Correct Rate (C) (%)	Rate of Common Mispronunciations (M)
/er/	4 755	889	19	**/ax/** (44%), /ee/ (16%)
/aa/	7 295	3 545	49	/ao/ (38%), **/ax/** (5%)
/ax/	13 734	7 498	55	/_/ (13%), **/ix/** (12%), **/ux/** (8%)
/ih/	8 578	5 054	59	**/ix/** (28%), /iy/ (7%)
/uh/	1 197	829	69	/uw/ (23%), **/ux/** (7%)
/eh/	4 320	3 054	71	**/ae/** (17%), /ih/ (3%)
/ae/	6 593	4 721	72	**/aa/** (14%), **/ax/** (8%), /eh/ (4%)

Note: Phonemes in boldface exist in American English but not in Chinese.

Table 4. English Consonants with the Lowest Rates of Correct Pronunciations C Based on the CU-CHLOE Corpus

Phoneme	Total Number of Pronounciations	Number of Correct Pronounciations	Correct Rate (C) (%)	Rate of Common Mispronunciations (M)
/r/	10 785	5 516	54	/_/ (38%), /w/ (6%)
/z/	4 400	2 551	58	/s/ (38%), /_/ (3%)
/th/	1 200	702	59	/f/ (38%), /_/ (2%)
/jh/	499	299	60	/_/ (13%), **/ch/** (12%), **/zh/** (4%), **/sh/** (4%)
/v/	3 100	1 899	61	/f/ (33%), /w/ (4%)
/d/	8 197	5 452	67	/_/ (15%), **/t/** (11%)
/dh/	6 099	4 189	69	/d/ (24%), /_/ (3%)

Note: Phonemes in boldface exist in American English but not in Chinese.

Target Pronunciation	Pronunciation by Native Chinese Speakers Speaking English																					
		er	aa	ax	ih	uh	eh	ae	iy	ey	ah	ow	uw	aw	ay	ao	oy	_	ux	ix	ee	axr	Total
	er	19%		44%									1%					6%			16%	10%	4 755
	aa		49%	5%							3%	2%				38%							7 295
	ax		1%	55%	1%			1%		1%		2%	2%			3%		13%	8%	12%			13 734
	ih				59%				7%						3%					28%			8 578
	uh					69%							23%						7%				1 197
	eh			2%	3%		71%	17%	2%											2%			4 320
	ae		14%	8%			4%	72%															6 593
	iy			3%	7%				78%	2%					1%			2%		7%			5 791
	ey		2%	4%	5%		8%	3%		76%													5 378
	ah		1%	7%							77%		2%	1%		9%							3 589
	ow										1%	82%	2%	1%		13%							3 099
	uw					5%						3%	83%					1%	7%				3 290
	aw		3%								3%	7%		84%		2%							899
	ay		1%	5%					1%					1%	86%					2%			4 799
	ao			2%							1%	2%				92%							4 299
	oy											5%				1%	93%						1 200

1% ~ 5% 5% ~ 20% 20% ~ 50% 50% ~ 100%

Fig.1. Mispronunciation matrix of English vowels by native Chinese speakers speaking English. Phonemes that are present in English but absent in Chinese are highlighted in yellow.

3.3 Discussions

In this subsection, a mispronunciation matrix of the English phoneme by native Chinese speakers is presented. If the phonemes are present in both English and Chinese, Chinese speakers tend to pronounce them correctly. On the contrary, if the phonemes are present only in English but not in Chinese, Chinese speakers tend to mispronounce them as other similar phonemes[11]. This phenomenon happens for both English consonants and vowels.

Inspired by this phenomenon, we further investigate how to find "similar phonemes" for a given English phoneme. In next section, the computational methods of obtaining the perceptual distance of vowels and consonants will be discussed.

4 Computational Methods of Perceptual Distance

Our approach considers that the speakers' perceptual space should be trained and formed based on their native language (L1). Therefore, in this section, we discuss the computational methods for perceptual distance (PD) of vowels and consonants, and apply the proposed methods to compute the auditory confusion of Chinese speakers.

We decompose this research question into two parts. First, we explore how acoustic features distinguish different vowels and consonants. Second, we verify the correlation between auditory perception and selected acoustic features of vowels and consonants. We devise four steps in building the perceptual space.

Step 1. For Chinese vowels and consonants, we choose the frequencies of 500 Hz, 1 000 Hz, and 2 000 Hz for this whole study, because these three frequencies are universally acknowledged as language frequencies. In audiology, language frequencies play a pivotal role for the perception of speech signals.

Step 2. The acoustic features which are usually used for distinguishing different vowels and consonants (such as speech recognition) are extracted.

Step 3. We define how to compute the distance of two vowels or consonants' acoustic features in mathematics.

Step 4. We conduct auditory experiments to record which vowels or consonants tend to be misheard, and obtain auditory confusion rate between two vowels or consonants. We continue to select the acoustic features from those extracted in step 2, until the trend of their distances which are defined in step 3 is the most consistent with the auditory confusion rates.

Then the distances computed in step 4 are called perceptual distance (PD) of Chinese vowels and consonants. In the next two subsections, we will introduce the feature selection, distance definition, and assessment of PD between vowels and consonants in details.

4.1 Computation of PD Between Vowels

4.1.1 *Feature Selection*

First of all, we need to choose a signal representation for computing the PD of vowels. We made a comparison between two commonly used spectrum modes: linear predictive coding (LPC) and cepstrum (CEP).

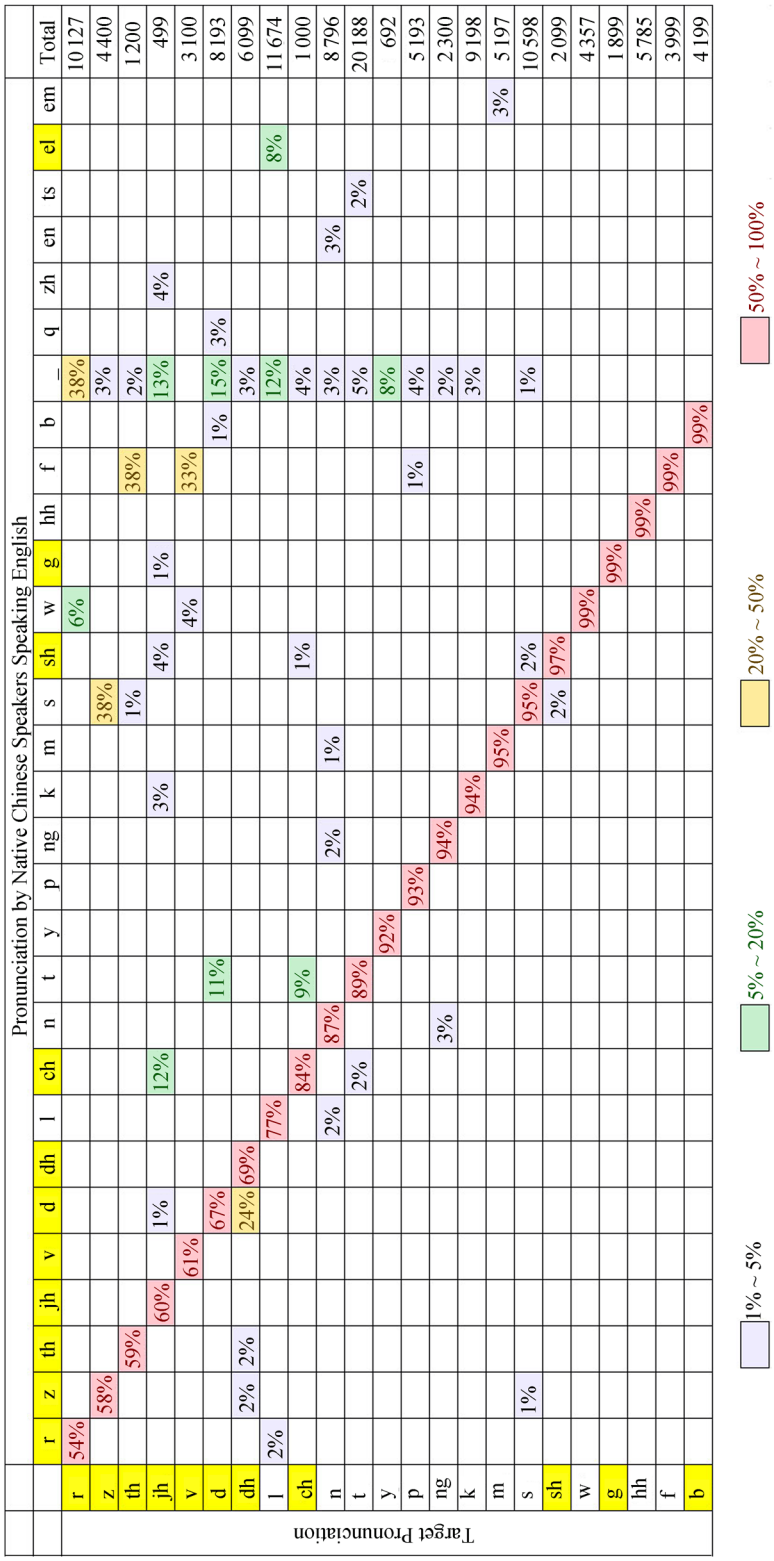


Fig.2. Mispronunciation matrix of English consonants by native Chinese speakers speaking English. Phonemes that are present in English but absent in Chinese are highlighted in yellow.

Fig.3 shows an example of the audio signal of the Chinese syllable "Ch**uang**". "Ch**uang**" with tone 1 and "Ch**uang**" with tone 2 are analyzed by using LPC and CEP respectively. LPC appears to be much better than

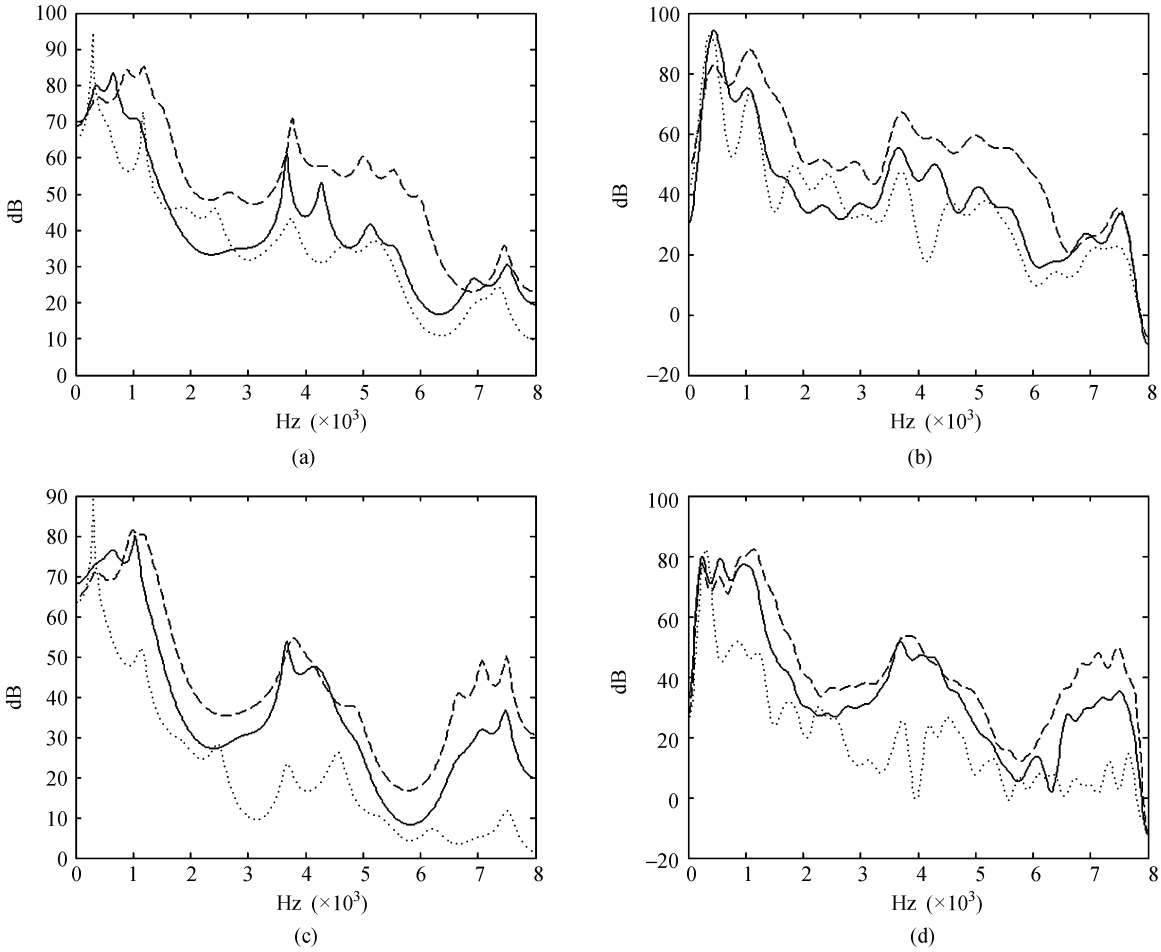


Fig.3. Comparison between LPC and CEP. The solid, dash and dotted lines indicate the spectrum of different frames respectively. (a) Spectrum of syllable "Ch**uang**" with tone 1 analyzed by using LPC. (b) Spectrum of syllable "Ch**uang**" with tone 1 analyzed by using CEP. (c) Spectrum of syllable "Ch**uang**" with tone 2 analyzed by using LPC. (d) Spectrum of syllable "Ch**uang**" with tone 2 analyzed by using CEP.

CEP in the analysis of frequency spectrum, in terms of clarity and regularity. Furthermore, the difference of spectrum between different tones is smaller by LPC than that by CEP. Hence, LPC is selected for analysis.

For each phoneme, we select three frames of the speech signal at 1/6, 3/6, 5/6 of the phoneme's duration. We resample the three frames (to 16KHz), window and reemphasize the signals to compute the LPC coefficients. We mathematically fit these coefficients to three spectral curves. Then, we can get values at 500, 1 000, 2 000 from the three curves by interpolation. Taking 500 Hz, 1 000 Hz, 2 000 Hz as the center respectively, we compute the integral of the LPC coefficients at frequency bands of [450, 550] Hz, [950, 1050] Hz, and [1950, 2050] Hz.

Since we have three spectral curves and three frequency bands, we can obtain a group of 9-dimensional vectors as the acoustic features of Chinese vowels. Furthermore, we define the perceptual distance (PD) between two vowels as the Euclidean distance of their feature vectors.

4.1.2 Assessment of PD of Vowels by Auditory Perception

In order to assess PD, we compare it with the raw acoustic distance between vowel classes. We put forward the concept of hierarchical clustering, and cluster all the vowel samples using PD. From the results of hierarchical clustering, we can figure out whether vowels are divided into the right classes or not. We set rules of hierarchical clustering as follows:

1) establish a corpus of Chinese vowels which are segmented from syllables pronounced by different speakers;

2) set every sample of vowels as a class;

3) cluster two classes to one between which the PD is the shortest in the set;

4) repeat last step until the number of classes is equal to the number of Chinese vowel categories.

The clustering results show that 74% of vowel samples belonging to the same vowel category are clustered to the same class on average, which indicates that the proposed PD can distinguish different vowels well.

We further conduct an auditory experiment to record which vowels tend to be misheard. In this way, we obtain auditory confusion rate between two vowels. We compare the PD (which is normalized to the interval [0, 1]) with the auditory confusion rate. Some examples are shown in Table 5.

Table 5. Comparison Result of Perception Distance and Auditory Confusion Rate

Confusion Pairs of Chinese Vowels		Auditory Confusion Rate	Perceptual Distance
Vowel A	Vowel B		
/in/	/iŋ/	0.152	0.394 2
/u͡aI͡/	/u͡än/	0.100	0.668 4
/ən/	/u͡ən/	0.077	0.701 4
/u͡än/	/vän/	0.073	0.731 0
/i͡ä/	/i͡ɑŋ/	0.056	0.954 8
/ən/	/ɤŋ/	0.053	1.000 0

The correlation index between the auditory confusion rate and the perceptual distance is about −0.6. In Table 5, we find that the closer the PD, the smaller the difference of an auditory perception. This indicates that PD can represent the auditory difference between two vowels well.

In the same way, we extract the LPC features to calculate the PD for vowels in the CU-CHLOE corpus. Take the vowels /er/ and /ey/ as examples. As Table 6 shows, the smaller the PD, the higher the mispronunciation rate. The detailed investigation on the correlation between PD and mispronunciation will be discussed in Section 5.

Table 6. Rates of Pronunciations and Perceptual Distances for /er/ and /ey/

Phoneme	Pronounced as	Rate of Pronunciation (%)	Perceptual Distance (PD)
/er/	/er/	17.4	0.000 0
/er/	/ax/	40.7	0.514 5
/er/	/axr/	9.6	0.598 9
/er/	/eh/	6.6	0.678 3
/ey/	/ey/	75.6	0.000 0
/ey/	/eh/	8.5	0.466 6
/ey/	/ih/	5.1	0.538 4

4.2 Computation of PD Between Consonants

4.2.1 Feature Selection

According to the perception definition computing method of Mandarin consonants[10], we extract 17 features, including average zero-crossing rate (1 dimension), MFCC coefficients (6 dimensions), and bark features (10 dimensions). The specific steps in extraction are:

1) Extract the zero-crossing rate according to the formula:

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

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 i-th sample point in the n-th frame;

2) Extract the mel frequency cepstral coefficient (MFCC), and select six of them which have been revealed to have strong correlation with consonant perception[10]: M1, M3, M4, M5, M8, and M11;

3) Extract the bark features, which include:

• calculating the FFT power spectrum,

• calculating the integral of the FFT power spectrum for each of the bark bands, marked as $x_1, x_2, \ldots, x_{21}$,

• calculating the bark rate y_i:

$$y_i = x_i \Big/ \sum_{j=1}^{21} x_j,$$

• selecting 10 bark rate features which have been revealed to have strong correlation with consonant perception[10]: B9, B10, B12, B16, B13, B19, B20, B21, B14, and B18;

4) Calculate the Euclidean distances between consonant feature vectors.

We use (1) to normalize the features, where f_{new} is the normalized feature, f_{source} is the feature to be normalized, and F_{source} is the set of original features.

$$f_{\mathrm{new}} = \frac{f_{\mathrm{source}} - \min(F_{\mathrm{source}})}{\max(F_{\mathrm{source}}) - \min(F_{\mathrm{source}})}. \tag{1}$$

Like vowels, we the define perceptual distance of consonants as the Euclidean distance of their feature vectors.

4.2.2 Assessment of PD of Consonants by Auditory Perception

We assess the PD of consonants as we did for vowels, based on the hierarchical clustering. The clustering results are shown in Fig.4.

From Fig.4, we can see that unvoiced and voiced consonants are separated at the first split. Voiced consonants (/m/, /n/, /l/ and /r/) are in the same cluster; radical consonants (/g/, /k/ and /h/) are in the same cluster; and supradental consonants (/ts/, /ts‘/, and /s/) are in the same cluster. The consonants with the

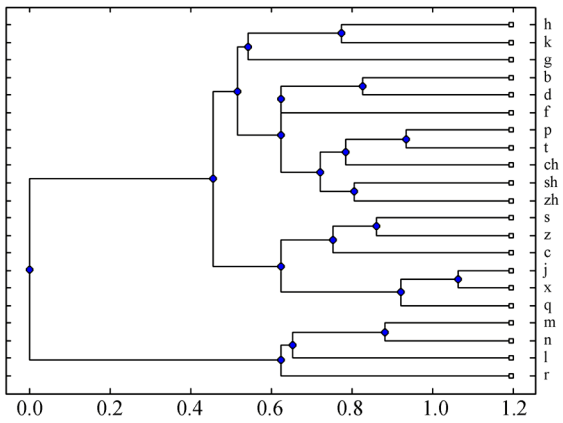


Fig.4. Results of hierarchical clustering of consonants by PD.

same place of articulation are also divided in the same cluster such as /tɕ/, /tɕ‘/ and /ɕ/. These results of clustering are quite consistent with the theory of phonetics[11].

We calculate the PD for consonants in training corpus of Enunciate system. As an example, we show the PD for the consonant /r/ in Table 7. The smaller the perceptual distance, the higher the mispronunciation rate. Hence, PD seems to relate with perception and pronunciation.

Table 7. Rates of Correct Pronunciation and Perceptual Distances for /r/

Phoneme	Pronounced as	Rate of Pronunciation (%)	Perceptual Distance (PD)
/r/	/r/	51.1	0.000 0
/r/	-	35.8	0.404 7
/r/	/ax/	6.1	0.468 6
/r/	/w/	5.7	0.523 3

5 Experiments on Correlation Analysis Between Mispronunciation Statistics and PD

Next, we experimentally investigate the correlation between mispronunciation statistics and PD.

5.1 Analysis of Phonemes with High Rates of Correct Pronunciation

First, considering the vowels and the consonants with the highest correct pronunciation 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:

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

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), and $dist(x_i, x_j)$ represents the perceptual distance between x_i and x_j.

As shown in Tables 8 and 9, the average perceptual distances of phonemes with the highest rate of correct pronunciation are quite high (compared with the results in Tables 10 and 11). The correlation between APD and C of vowels with the highest rates of correct pronunciation is 0.73, while the correlation between APD and C of consonants with the highest rates of correct pronunciation is 0.86. In general, the APD and the correct rate C are in direct proportion (Figs. 5 and 6). It indicates that 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.

Table 8. Average Perceptual Distances of Vowels with the Highest Rates of Correct Pronunciation

Phoneme	Correct Rate (C) (%)	Average Perceptual Distance (APD)
/oy/	93.1	0.903 5
/ao/	91.5	0.914 6
/ay/	86.4	0.880 3
/aw/	84.2	0.852 6
/uw/	82.3	0.863 1
/ow/	81.6	0.740 6
/ah/	77.0	0.736 7

Table 9. Average Perceptual Distances of Consonants with the Highest Rates of Correct Pronunciation

Phoneme	Correct Rate (C) (%)	Average Perceptual Distance (APD)
/b/	99.2	0.914 0
/f/	99.1	0.924 6
/hh/	99.1	0.892 3
/g/	98.8	0.852 6
/w/	97.7	0.863 1
/sh/	96.8	0.840 6
/s/	95.1	0.846 7

Table 10. APD of 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%)	0.513 7
/aa/	48.6	/ao/ (38.3%), /ax/ (5.4%)	0.556 7
/ax/	54.0	/-/ (12.6%), /ix/ (11.5%), /ux/ (7.5%)	0.482 9
/ih/	58.1	/ix/ (27.9%), /iy/ (6.6%)	0.4911
/uh/	69.1	/uw/ (23.3%), /ux/ (6.8%)	0.666 7
/eh/	69.4	/ae/ (17.1%), /ih/ (3.4%)	0.745 0
/ae/	71.6	/aa/ (13.7%), /ax/(8.2%), /eh/ (4.2%)	0.684 9

Table 11. *APD* of Consonants with the Lowest Rates of Correct Pronunciation

Phoneme	Correct Rate (C) (%)	Rate of Common Mispronunciations (M)	Average Perceptual Distance (APD)
/r/	51.1	/_/ (35.8%), /w/ (5.7%)	0.464 0
/z/	58.0	/s/ (38.3%), /_/ (2.9%)	0.563 7
/th/	58.5	/f/ (37.7%), /_/ (1.6%)	0.512 4
/jh/	59.8	/_/ (13.0%), /ch/ (12.0%), /zh/ (4.4%), /sh/ (3.6%)	0.514 5
/v/	61.3	/f/ (32.7%), /w/ (3.9%)	0.413 0
/d/	66.5	/_/ (15.3%), /t/ (10.7%)	0.630 4
/dh/	68.7	/d/ (23.6%), /_/ (2.8%)	0.556 0

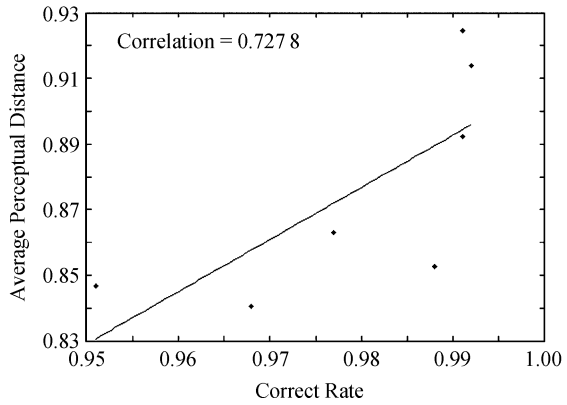


Fig.5. Correlation between APD and C of vowels with the highest rates of correct pronunciation.

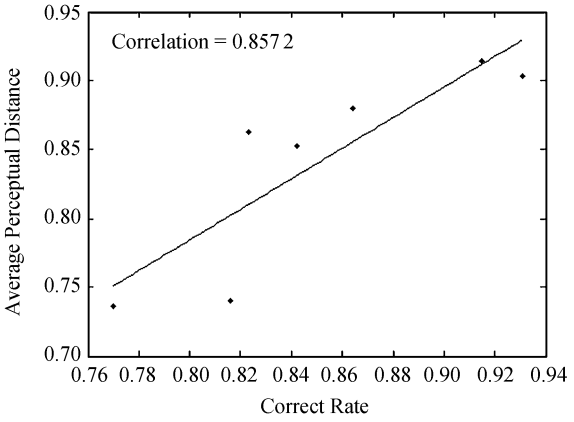


Fig.6. Average perceptual distances of consonants with the highest rates of correct pronunciation.

5.2 Analysis of Phonemes with Low Rates of Correct Pronunciation

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

$$APD(/er/) = 1/2(dist(/er/, /ax/) + dist(/er/, /ee/)).$$

As shown in Tables 10 and 11, APD and correct rate C are positively correlated (Figs. 7 and 8). The correlation between APD and C of vowels with the lowest rates of correct pronunciation is 0.65, while the correlation between APD and C of consonants with the lowest rates of correct pronunciation is 0.53. In general, 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 3.

In summary, APD and the correct rate C are in direct proportion, while the mispronunciation rate M is negatively correlated with auditory perception distance APD. It indicates that the auditory confusion indirectly reflects the spoken confusion.

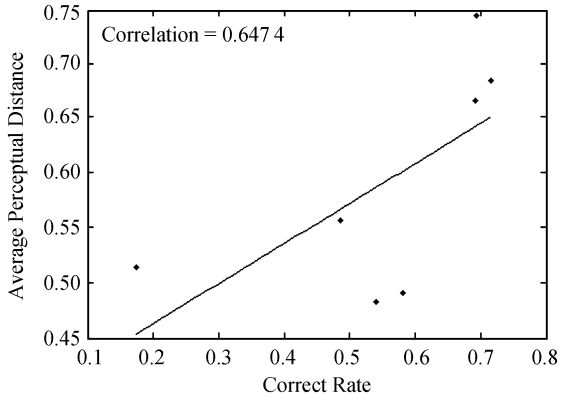


Fig.7. Average perceptual distances of vowels with the lowest rates of correct pronunciation.

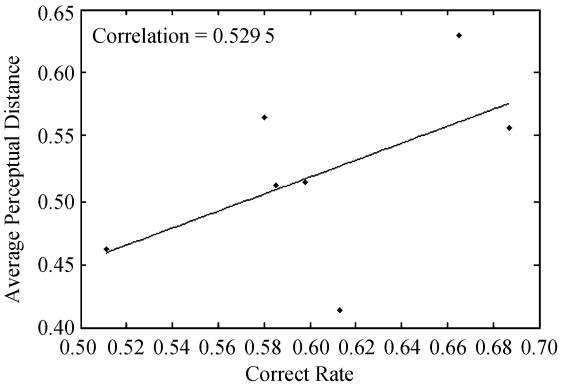


Fig.8. Average perceptual distances of consonants with the lowest rates of correct pronunciation.

6 Gradation of Mispronunciations in L2 (English) Learning

In order to further facilitate L2 (English) learning, we aim to further propose principles for generating a gradation of the severity of mispronunciations. Since

our study targets Chinese (L1) speakers learning English (L2), we attempt to propose several directions on gradation of L2 English mispronunciations. Our ultimate goal is to use the proposed principles for pedagogically improving prioritization corrective feedback of CAPT systems.

Previous methodologists on L2 learning usually suggest that teachers focus their attention on a few mispronunciation error types rather than try to address all the errors[7-8]. If too many errors are presented to learners at the same time, learners may get confused and discouraged. Therefore, the PD proposed in this paper can be used to help us grade the severity of errors[12].

6.1 Gradation Based on Computational Perceptual Relevance

From the previous sections, we find that the frequency of an error (phone α is mispronounced as phone β) reduces with the increase in perceptual distance (PD) between α and β. PD can be used to some extent as a measure of how easily learners tend to confuse certain phonemes. The smaller the PD between two phonemes, the more likely they will be confused by L2 learners, the higher error frequency we can observe from the results. Therefore, the gradation of errors in terms of severity is negatively correlated with the observed error frequency but positively correlated with PD.

6.2 Gradation Example

Let us finally give a gradation example. Based on the above CU-CHLOE corpus used in Section 3, we first compute the PD of every pair of vowels and consonants. Then the pairs of vowels and consonants with lower PD values are selected. We further validate these phoneme pairs with the statistic results of mispronunciation. Finally, 10 pairs of phonemes with the lowest PD values as well as the highest mispronunciation rates are chosen as the highest-priority order of corrective feedback generated for the L2 learners.

So the final results are: for Chinese speakers to learn English, we suggest that learners need to pay more attention to the following phonemes:

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

where "A => $B, C, \ldots$" means A is most easily to be mispronounced as B or C or ....

7 Conclusions

In this paper, we presented a computational approach of obtaining the auditory perceptual distance between phonemes, and proposed to utilize it for quantitatively representing a gradation of pronunciation errors in L2 English speech by L1 Chinese learners. The correlation between mispronunciation statistics and perceptual distances was experimentally investigated. This correlation finally led to some suggestions on the prioritization of corrective feedback generation for CAPT.

The main conclusions of this paper are:

1) The auditory perceptual distance and the correct pronunciation rate are positively correlated;

2) PD can be used as a systematic way of grading mispronunciations and prioritizing them for corrective feedback generation.

Future work will focus on how to deal with the insertion or deletion in mispronunciation.

References

[1] Braj K. Asian Englishes: Beyond the Canon. Hong Kong: Hong Kong University Press, 2005.

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

[3] Meng H, Lo Y, Wang L, Lau W Y. Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, December 2007, pp.437-442.

[4] Lo W K, Harrison A M, Meng H, Wang L. Decision fusion for improving mispronunciation detection using language transfer knowledge and phoneme-dependent pronunciation scoring. In *Proc. the 6th International Symposium on Chinese Spoken Language Processing*, December 2008, pp.25-28.

[5] Yuen K W, Leung W K, Liu P F, Wong K H, Qian X, Lo W K, Meng H. Enunciate: An internet-accessible computer-aided pronunciation training system and related user evaluations. In *Proc. International Conference on Speech Databases and Assessment*, October 2011, pp.85-90.

[6] Laver J. Principles of Phonetics. Cambridge, UK: Cambridge University Press, 1994.

[7] Ellis R. Corrective feedback and teacher development. *L2 Journal*, 2009, 1: 3-18.

[8] Wang H, Qian X, Meng H. Phonological modeling of mispronunciation gradations in L2 English speech of L1 Chinese learners. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, May 2014.

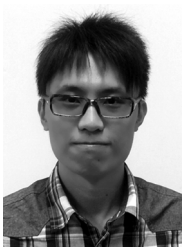
[9] Huang G, Jia J, Cai L. A study on perception measurement of mandarin vowels based on LPC spectrum features. In *Proc. Phonetic Conference*, May 2010.

[10] Jia J, Wang Y, Zhang Y, Tian Y, Cai L. Discussion on perception definition computing method of mandarin consonants. In *Proc. Phonetic Conference*, May 2012.

[11] Meng H, Zee E, Lee W S. A contrastive phonetic study between Cantonese and English to predict salient mispronunciations by Cantonese learners of English. Technical Report, SEEM2007-1500, Department of Systems Engineering and

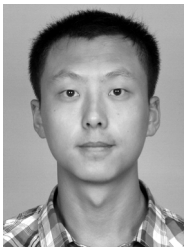
Engineering Management, the Chinese University of Hong Kong, February 2007.
[12] Neri A, Cucchiarini C, Strik H, Boves L. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 2002, 15(5): 441-467.

Jia Jia is an associate professor in the Department of Computer Science and Technology, Tsinghua University, Beijing. She got her B.S. and Ph.D. degrees both in computer science and technology from Tsinghua University in 2003 and 2008 respectively. Her main research interest is human computer speech interaction and social affective computing.

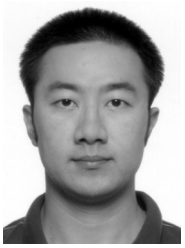


Wai-Kim Leung received his B.E. degree in systems engineering and engineering management from The Chinese University of Hong Kong in 2010. He is a master student in the Department of Computer Science and Technology at Tsinghua University from 2014. Before that, he served as a research assistant at the Human-Computer Communications Laboratory and the Department of Systems Engineering and Engineering Management at The Chinese University of Hong Kong. He focuses on system design and development of computer-aided pronunciation system.

Yu-Hao Wu received her B.E. degree in computer science and technology from Tsinghua University in 2012. She is currently a Ph.D. student at the Research Group of Human Computer Speech Interaction, Department of Computer Science and Technology at Tsinghua University. She focuses on computational methods of speech perceptual distance.



Xiu-Long Zhang received his B.E. degree in precision instruments and mechanology from Tsinghua University in 2009. He is currently a master student in Department of Computer Science and Technology of Tsinghua University. His research interests include pure tone audiometry and Mandarin speech audiometry.



Hao Wang received his B.E. degree in systems engineering and engineering management from The Chinese University of Hong Kong in 2010. He is currently a Ph.D. candidate in the Department of Systems Engineering and Engineering Management of the university with research interests primarily in nonnative speech and language processing for computer-aided language learning. His work focuses on gradation of mispronunciations in nonnative speech according to different levels of severity.

Lian-Hong Cai is a professor in the Department of Computer Science and Technology, Tsinghua University. She is now in charge of the Research Group of Human Computer Speech Interaction of the department. She received her B.E. degree in computer science and technology from Tsinghua University in 1970. Her main research interest is human computer speech interaction.

Helen M. Meng received all her degrees in electrical engineering from MIT. She is a professor and chairman of the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. She is also the founding director of the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies, which has been recognized as a Ministry of Education of China (MoE) Key Laboratory since 2008. She is elected into the IEEE Board of Governors in 2014. And she is a fellow of the Hong Kong Computer Society, Hong Kong Institute of Engineers and was also elevated to IEEE Fellow in 2013 for her contributions to spoken language and multimodal systems.