Developing a Computer-Aided Pronunciation System for Chinese-Speaking Learners of English
面向中国英语学习者的计算机辅助发音系统

Helen MENG, Alissa HARRISON and Lan WANG

ABSTRACT In this paper, we summarize the predicted phonetic confusions from our comparative analysis of Cantonese and English. Then, we describe the context-sensitive rules used to generate variants in the extended pronunciation lexicon. Finally we present some experimental results using learner data from our CU-CHLOE corpus to demonstrate the accuracy of our system.

KEYWORDS computer-aided pronunciation; comparative analysis; context-sensitive rules

摘　要　本文通过对比分析粤语和英语，总结了可预测的音素混淆规则。同时，描述了使用上下文相关的规则来产生扩展发音词典里的各种情况。最后，使用学习者数据库CU-CHLOE进行实验，通过实验结果证实了本系统的准确性。

关键词　计算机辅助发音训练；比较分析；上下文相关规则

I　Introduction

In this paper we summarize our work in developing a computer-aided pronunciation training (CAPT) system for Chinese-speaking learners of English. Our system is grounded in the theory of language transfer: knowledge of the first language (L1) sound system is carried over to the second language (L2). From a comparative phonological analysis of Chinese (Cantonese) and English, we develop a set of predicted phonetic confusions. These confusions – formulated as context-sensitive rules – are used to generate a pronunciation lexicon extended with common mispronunciations. This pronunciation lexicon is then used with an HMM-based speech recognizer to detect and diagnose salient segmental mispronunciations in Chinese-speaking learners of English. This diagnostic feature is unique to this system and is not seen in previous pronunciation scoring methods[1-3].

Our system is designed to be a supplemental resource to an existing English teaching curriculum. The target audience of the system is advanced adult learners (high school or university students) who are native Chinese speakers. It can benefit students by providing an always-accessible intelligent tutor for English speaking practice. This is particularly useful for environments where the teacher-to-student ratio is high or where there are otherwise few opportunities for speaking English.

In this paper, we summarize the predicted phonetic confusions from our comparative analysis of Cantonese and English. Then, we describe the context-sensitive rules used to generate variants in the extended pronunciation lexicon. Finally we present some experimental results using learner data from our CU-CHLOE corpus to demonstrate the accuracy of our system.

2　Comparative Analysis

This section is an abbreviated summary of our comparative analysis between Cantonese and English. A more detailed description can be found in [4] and empirical validation of these predicted confusions is available in [5].

2.1 Vowels

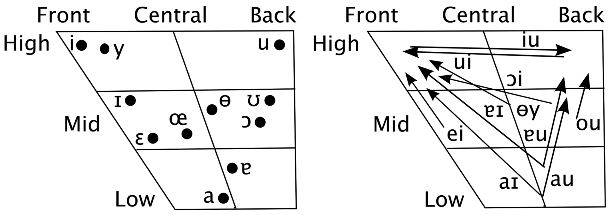

Figure 1 Cantonese vowels, adapted from [7]

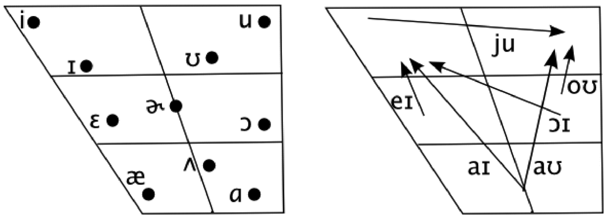

Figure 2 English vowels, adapted from [8]

The Cantonese vowel inventory is generally richer than English, however it lacks the rhotic and many low vowels found in English. Figure 1 illustrates the monophthongs and diphthongs found in Cantonese, and Figure 2 illustrates the English vowels. A comparison of the two vowel inventories shows that the English vowels [æ], [ɚ], [ʌ], and [ɑ] are missing in Cantonese. The lack of these vowels in the learner's L1 is hypothesized to be a source of possible mispronunciation in the L2 [6]. Furthermore, learners are predicted to produce sounds from their L1 which are phonetically-similar to the L2. Some examples of these phonetic confusions for vowels include 'had' [hæd] produced as 'head' [hɛd] and 'her' [hɚ] produced as [hœ].

2.2 Consonants

There are several consonants in English which are not found in Cantonese. Table 1 and Table 2 show the consonants in the two languages. The English consonants missing in Cantonese can be generally be grouped into the following classes: (1) voiced stops [b], [d], and [g]; (2) post-alveolar affricates [tʃ] and [dʒ]; (3) interdental fricatives [θ] and [ð]; (3) voiced fricatives [v], [z], [ʃ], and [ʒ]; and (4) retroflex approximant [r].

Voiced stops in English are devoiced syllable-initially. This makes them phonetically equivalent to the unaspirated stops of Cantonese and not problematic for learners in syllable-initial contexts. However, elsewhere the English voiced stops retain voicing and are often substituted with the unaspirated counterpart, e.g. the word 'cab' [kæb] is confused with 'cap' [kæp].

The post-alveolar affricates [tʃ] and [dʒ] are often substituted with the alveolar affricates [ts^h] and [ts], respectively.

The voiceless interdental fricative [θ] is typically replaced with the labiodental fricative [f]. This can been seen in the word 'three' [θri] being confused with 'free' [fri]. The voiced interdental fricative [ð] is often mispronounced as [t], e.g. 'there' [ðɛr] confused with 'dare' [dɛr].

The remaining voiced fricatives are typically produced with the voiceless counterpart. For example, the word 'seize' [siz] is confused with 'sees' [sis]. The post-alveolar fricatives are both produced as the voiceless alveolar fricative, e.g. 'show' [ʃoʊ] is confused with 'so' [soʊ].

The English retroflex approximant [r] should be produced with both rounded lips and retroflexion. It does not exist in Cantonese and learners tend to substitute other similar approximants, such as [w] or [l]. Examples of this are seen in the confusion of the words 'rate' [reɪt], 'wait' [weɪt] , and 'late' [leɪt].

2.3 Phonotactic Constraints

The syllable structure of Cantonese is remarkably more restrictive than English. Cantonese does not have consonants clusters. Furthermore, it has a highly-restricted set of possible coda consonants ([m], [n], [ŋ], [p], [t], [k]). English, on the other hand, can have clusters of up to three consonants in both the onset and coda positions, e.g. [strɛŋθs]. Due to the restrictive phonotactic constraints of Cantonese, learners often delete consonants or insert vowels to simplify the syllable structure. For example, the word 'professor' [prəfɛsɚ] becomes [poʊfɛsa] due to consonant deletion. Alternatively, the word 'kissed' [kɪst] may become two-syllables due to vowel insertion [kɪstə].

Table 1 Cantonese consonants, adapted from [7]

	Labial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosivel	p 爸 p^h 爬		t 打 t^h 他			k 加 k^h 加 k^w 加 k^{wh} 侉	
Affricates			ts 抓 ts^h 叉				
Nasals	m 妈		n 拿			ŋ 牙	
Fricatives	f 花		s 沙				h 虾
Approxi-mation					j 忧	w 蛙	
Laterals			l 拉				

Table 2 English consonants, adapted from [8]

	Labial	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosivel	p_{pie} b_{buy}		t_{tie} d_{die}			k_{kite} g_{guy}	
Affricates				tʃ_{chin} dʒ_{gin}			
Nasals	m_{my}		n_{nigh}			ŋ_{hang}	
Fricatives	f_{fie} v_{vie}	θ_{thigh} ð_{thy}	s_{sigh} z_{zoo}	ʃ_{shy} ʒ_{azure}			h_{high}
Approxi-mation				ɹ_{rye}	j_{you}	w_{why}	
Laterals			l_{lie}				

3 Variant Generation

In predicting phonetic confusions, it is not sufficient to use simple one-to-one mapping rules without reference to the phonetic environment [9]. These type of mappings, also known as context-insensitive rules, lead to a significant over-generation of pronunciation variants for a word. For example, we know that /d/ may be deleted and this can be represented as /d/ →∅. For the word "did" /d ih d/, this single mapping generates four possible variants including /ih/. However, this variant is known to be highly implausible in Chinese-speaking learners' English.

Context-sensitive rules can effectively solve the problem of variant over-generation. These rules specify a phonetic environment which must be satisfied in order for the rule to apply. This condition is conventionally denoted by a forward slash following the mapping.

Additionally, phonetic environments are described using convenient abbreviations of various linguistic classes: C for consonants, word-boundaries. As noted in the discussion of phonotactic constraints, many of the predicted phonetic confusions are constrained to particular environments. In particular, due to the lack of final voiced stops and consonants clusters in Cantonese, we can state that the deletion of /d/ is constrained to word-final positions or before another consonant. This is represented in context-sensitive rules as: /d/ → ∅ / _ # and /d/ → ∅ / C _ .

The use of context-sensitive rules can reduce the number of generated pronunciation variants by nearly an order of magnitude. Moreover, since the context-sensitive rules can characterize the phonetic confusions made by learners, the variants generated are far more plausible. This has been demonstrated by increased accuracy of the detection and diagnosis of mispronunciations in [9].

4 Corpus Development

As part of this research initiative, we have also established a corpus of learner data for system development and evaluation. The Chinese University CHinese English Learners Of English (CU-CHLOE) corpus has recordings of read English from over 200 native speakers of Chinese. The corpus includes three types of prompts: (1) The North Wind and the Sun; (2) minimal pairs, confusable word groups, and difficult sentences selected by English teachers in our university; (3) sentences from the TIMIT database [10]. These prompts were selected to ensure a balance of phonetic environments and representative examples of mispronunciations from Chinese-speaking learners. The recordings were all carried out in a sound-dampened room with a high-quality noise-canceling headset (Sennheiser PC155). In the experimental results of this paper, we have utilized a pilot collection of the CU-CHLOE corpus. This pilot collection includes 21 native Cantonese speakers reading the short passage. The North Wind and the Sun which has been annotated by an expert human listener.

5 Experimental Results

5.1 Speech Recognizer

The system uses a cross-word triphone HMM-based speech recognizer with the HTK Toolkit [11]. Each HMM has three states and each state has 12 Gaussian mixtures. There are a total of 39 features: PLP + Δ + $\Delta\Delta$ with cepstral mean normalization. Altogether, there are 688 unique HMM states and 1987 unique models after state-tying. The training data comes from the TIMIT training set which contains a total of 4620 sentences recorded by 462 speakers representing all the major dialects of the U.S. The recognizer is run in force-alignment mode using the word-level transcription of the prompt and the extended pronunciation lexicon. The output of the recognizer is a phone-level transcription of the learner's utterance.

5.2 Evaluation procedures

The recognized output is aligned with the transcription from the (1) human annotator and (2) the model pronunciation. This alignment can be done using a dynamic programming algorithm that minimizes the sum-of-pairs score between the three transcriptions [12]. Insertions and deletions both were given a cost of seven, while substitutions were weighted based on phonetic distance. An example of the three string alignment is given in Table 3.

Table 3 Example of three-string alignment

MODEL:	n	ao	r	th
GOLD STANDARD:	l	ao		th
SYSTEM:	n	aa		th

Mispronunciation detection is classified in four categories: true acceptance (TA), true rejection (TR), false acceptance (FA), and false rejection (FR). TA occurs when both the system and the human annotator label a phone as identical to the model pronunciation (/th/ in Table 3). When both the system and human annotator differ from the model, it is a TR (/r/ deletion in Table 3). FR occurs when the system differs from the model but the human annotator considers it identical, and vice versa with FA.

5.3 Mispronunciation Detection

The performance of the system mispronunciation detection ability is measured in terms of false acceptance rate (FAR) and false rejection rate (FRR). These measures are defined as:

$$FalseAcceptanceRate(FAR) = \frac{FA}{TR + FA}$$

$$FalseRejectionRate(FRR) = \frac{FR}{TA + FR}$$

In our experimental results, only 14.9% of phones were falsely rejected as mispronounced by the system (FRR) while 43.6% of mispronunciations were falsely accepted (FAR). Note that there is an inevitable tradeoff in the FAR and FRR. While the goal is to keep both at a minimum, we believe a lower FRR is preferable to the alternative. A lower FRR means that the system will be less likely to reject correct pronunciations at the

[1] In this and the following sections, we use the ASCII-based notation system known as DARPABET for phonetic transcription. These DARPABET transcriptions can be distinguished from IPA transcriptions by the use of slashes instead of brackets.

cost of failing to detect some mispronunciations. This is important as a learning tool should encourage the learner first and foremost.

5.4 Mispronunciation Diagnosis

The mispronunciation diagnosis of the system was measured by the calculating the percentage of phone labels which are identical between the system and the human annotator for the true mispronunciations of the learners. In diagnostic performance, the system correctly identified the phone in 51.0% of the mispronunciations (where there are 44 possible phone labels).

6 Conclusions

In this work, we have demonstrated a novel CAPT system design for Chinese-speaking learners of English. Our method utilizes linguistic-knowledge to create a system which can not only detect but also diagnose mispronunciations. We believe this approach can be applied to other language pairs and are currently working on incorporating the phonetic confusions of Mandarin into the system. In future work, we plan to improve the accuracy of our system using discriminative training techniques. We also plan to extend the system to the detection and diagnosis of suprasegmental errors.

7 Acknowledgments

This work is partially supported by the CUHK Teaching Development Grant. The authors would also like to acknowledge the contributions of Professor Eric Zee, Professor Wai-Sum Lee, Dr. Pauline Lee, Dr. Wai-Kit Lo, Patrick Chu, Wing Yiu Lau, and Yuen Yee Lo to this project.

REFERENCES

[1] J. Mostow, A. G. Hauptmann, L. L. Chase, *et al*. Towards a Reading Coach that Listens: Automated Detection of Oral Reading Errors. Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93), American Association for Artificial Intelligence, Washington, DC, July 1993, pp. 392-397.

[2] S. M. Witt , S. J. Young, Performance Measures For Phone-Level Pronunciation Teaching in CALL. Proc. of the Workshop on Speech Technology in Language Learning, pp. 99-102, Marholmen, Sweden, 1998.

[3] B. Mak, M. H. Siu, M. Ng, *et al*. PLASER: Pronunciation Learning via Automatic Speech Recognition. Proceedings of HLT-NAACL, 2003.

[4] H. Meng, E. Zee, W. S. Lee. Deriving Salient Learners' Mispronunciations from Cross-language Phonological Comparisons. CUHK Technical Report, SEEM2007-1500, February 2007.

[5] H. Meng, Y. Y. Lo, L. Wang, *et al.* Deriving Salient Learners' Mispronunciations from Cross-language Phonological Comparisons. in Proceedings of ASRU 2007.

[6] R. Lado. Linguistics Across Cultures: Applied Linguistics for Language Teachers. Uni of Michigan, Ann Arbor, 1957.

[7] Zee, E. Chinese (Hong Kong Cantonese). Journal of International Phonetic Association, 21:1, 1991.

[8] P. Ladefoged. American English. Handbook of the IPA. Cambridge University Press, 1999.

[9] A. M. Harrison, W. Y. Lau, H. Meng, L. Wang. Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. in Proceedings of Interspeech 2008.

[10] W. Fisher, V. Zue, J. Bernstein, *et al.* An Acoustic-Phonetic Data Base. J. Acoust. Soc. Am. 81, Suppl 1, 1987.

[11] S. J. Young, J. Odell, D. Ollason, *et al.* The HTK book Entropic. Cambridge Research Laboratory, 1996.

[12] D. Gusfield. Algorithms on strings, trees, and sequences. CUP, NY, 1997.

作者简介

Helen MENG ：作者简介见本期封2页。

Alissa HARRISON： A Research Assistant in the Human-Computer Communications Laboratory at the Chinese University of Hong Kong. She received her B.S. in Computer Science and B.A. in Linguistics at the University of Washington and her M.Phil. in Linguistics at the Chinese University of Hong Kong in 2008. Her research interests are in second language acquisition, speech recognition, and phonetics.

Lan WANG：作者简介见本期封2页。