DERIVING SALIENT LEARNERS' MISPRONUNCIATIONS FROM CROSS-LANGUAGE PHONOLOGICAL COMPARISONS

Helen Meng[1,2], Yuen Yee Lo[2], Lan Wang[2] and Wing Yiu Lau[1]
[1]Human-Computer Communications Laboratory, The Chinese University of Hong Kong (CUHK)
[2]Ambient Intelligence and Multimodal Systems Laboratory,
CAS/CUHK Shenzhen Institute of Advanced Integration Technologies, Chinese Academy of Sciences
{hmmeng, wylau}@se.cuhk.edu.hk, {wy.lu, lan.wang}@siat.ac.cn

ABSTRACT
This work aims to derive salient mispronunciations made by Chinese (L1 being Cantonese) learners of English (L2 being American English) in order to support the design of pedagogical and remedial instructions. Our approach is grounded on the theory of language transfer and involves systematic phonological comparison between two languages to predict *possible phonetic confusions* that may lead to mispronunciations. We collect a corpus of speech recordings from some 21 Cantonese learners of English. We develop an automatic speech recognizer by training cross-word triphone models based on the TIMIT corpus. We also develop an "extended" pronunciation lexicon that incorporates the predicted phonetic confusions to generate additional, erroneous pronunciation variants for each word. The extended pronunciation lexicon is used to produce a confusion network in recognition of the English speech recordings of Cantonese learners. We refer to the statistics of the erroneous recognition outputs to derive salient mispronunciations that stipulates the predictions based on phonological comparison.

Index Terms — Language learning, mispronunciation detection, phonetic and phonological analysis

1. INTRODUCTION

The objective of this work is to derive salient mispronunciations made by Cantonese (L1) learners of English (L2). Our long-term goal is to design effective pedagogical and remedial instructions for pronunciation improvement. The target learners are adults (high school and university students) who are native Cantonese speakers seeking to improve their pronunciation in English. We present a methodology that predicts possible phonetic confusions based on a comparative phonological analysis between L1 and L2. These predicted confusions are incorporated into a pronunciation lexicon to generate additional, erroneous pronunciation variants of each word. The extended pronunciation lexicon is used to produce a confusion network for recognition of English speech recordings of Cantonese learners. We tabulate the statistics of the erroneous recognition outputs to derive salient mispronunciations that stipulates the predictions based on phonological comparison.

Pronunciation errors may be due to a diversity of factors, such as an imperfect understanding of semantics, syntax, morphology, phonology, coarticulatory effects and letter-to-sound rules. As an initial step, we focus on phonology. Previous efforts have incorporated automatic speech recognition in computer-assisted language learning (CALL) and/or pronunciation measurement for non-native speakers [1-7]. Typically L1 and L2 comparisons are made between expert transcriptions of an available speech corpus and its canonical transcriptions based on the native model. Our current work is also grounded on the theory of language transfer, but differs in the sense that we conduct phonological comparisons between L1 and L2 across the phonetic and phonotactic levels to identify major disparities, such as missing phones and violation of phonotactic constraints, in order to focus on phonological contexts where perceived interferences of transfer features are prominent. The linguistic discrepancies may offer an explanatory model for us to target specific errors and understand their causes. Such understanding will be beneficial for the design of pedagogical and remedial instructions for pronunciation improvement.

The outline of this paper is as follows: Section 2 presents the comparative phonological analysis between L1 (Cantonese) and L2 (English). Section 3 describes the speech corpus that we are collecting to support diagnosis of mispronunciations. Section 4 details our experimental results and analysis. The conclusions and directions of future work is provided in Section 5.

2. PHONETIC COMPARISION

2.1. Vowels and Diphthongs

Figure 1 illustrates the Cantonese vowels charts of Cantonese containing 4 short vowels, 7 long vowels and 10 diphthongs. Chinese characters whose syllable pronunciations contain these vowels and diphthongs are listed in the Appendix A. Figure 2 illustrates the American English vowel charts containing 13 vowels and 3 diphthongs in American English. Appendix B presents English words containing these vowels and diphthongs [8]. The reduced vowel /ə/ is excluded because its quality varies considerably based on coarticulatory context. Comparison

between Figures 1 and 2 helps organize our observations on common mispronunciations due to English vowels that are *missing* in the Cantonese phonetic inventory. The missing set includes */e, æ, o, ɚ, ʌ, a/*. Hence, native Cantonese speakers often replace the missing English vowels with Cantonese vowels that are close in term of production and perception. Depending on the degree of resemblance, a subset of these vowels may be perceived as mispronunciations, due to prominent transfer effects from Cantonese (L1) to English (L2). Illustrative examples include pronouncing "had" */hæd/* as "head" */hɛd/;* or "her" */hɚ/* without the retroflexion as */h œ/.*

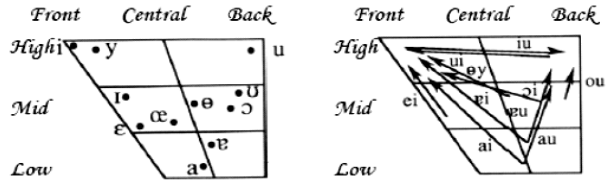

Figure 1. Cantonese vowels and diphthongs, based on [9]. Tongue positions (front, central, back, high, mid, low) are labeled.

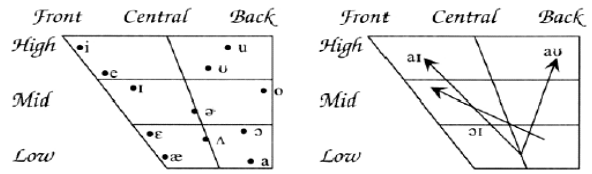

Figure 2. American English vowels & diphthongs, based on [10].

2.2. Consonants

Tables 1 and 2 show the consonants in Cantonese and American English respectively, organized according to the manner and place of articulation. Comparison between the two tables helps structure our observations in common mispronunciations for Cantonese learners of English. We refer specifically to English consonants that are *missing* from the Cantonese inventory, including voiced plosive, fricatives and affricates. Most Cantonese learners often substitute the missing English consonants with Cantonese consonants with similar place and/or manner of articulation. We present details in the following subsection.

2.2.1. Missing Voiced Plosives

The voiced plosives */b, d, g/* are present in English but absent in Cantonese. In the *prevocalic* position, these are often substituted with the voiceless, unaspirated Cantonese plosives */p, t, k/* which serve as good approximations. However, in the postvocalic position, voiced plosives may be unaspirated and voicing may be realized as durational lengthening of the preceding syllable nucleus. This leads to the durational difference within the contrastive word pairs: "cab" versus "cap" (*/k æ b/* vs. */k æ p/)*, or "pad" verse "pat" (*/p æd/* vs. */p æ t/).* These words pairs are often not clearly distinguished by Cantonese learners. Additional examples include "feed" */f i d/* is pronounced as "feet" */f i t/*; or "bag" */b æ g/* is pronounced as "back" */b æ k/*, etc.

2.2.2. Missing Affricates

English affricates are post-alveolar and include unvoiced and voiced tokens, namely, */tʃ, dʒ/.* These are non-existent in Cantonese and are often replaced respectively with the aspirated and unaspirated alveolar affricates */ts^h, ts/.* These have close resemblance in the place of articulation.

2.2.3. Missing Fricatives

This subsection addresses English fricatives that are missing from the Cantonese inventory. We describe the common substitutions performed by Cantonese learners of English. The English */v/,* a voiced, labiodental fricative, is often mispronounced either as the voiceless labiodental fricative /f/ or the sonorant bilabial approximant */w/,* as shown in the following examples*:*

- "vast" */v æ s t/* vs. "fast" */f æ s t/*
- "vest" */v ɛ s t/* vs. "west" */w ɛs t/*

There are two English dental fricatives. */θ/* is voiceless and is often mispronounced as the voiceless Cantonese labiodental */f/*. /ð/ is voiced and is often mispronounced as the voiced alveolar plosive in Cantonese */t/.* Examples include:

- "three" */θ r i* / versus "free" */f r i/*
- "there" */ð ɛ r/* versus "dare"*/d ɛ r/*

The English alveolar, voiced fricative /*z*/ is often mispronounced as the voiceless /*s*/, such as

- "seize" */s i z/* vs. "see" */s i s/*
- "zinc" */z ɪ ŋ k/* vs. "sink" */s ɪ ŋ k/*

The English pre-palatal (post-alveolar) /*ʃ ʒ*/ are frequently substituted with voiceless /s/, such as:

- "show" / *ʃ o/* vs. "so" */s o/*
- "social" */s o ʃ ə l/* vs. "soso" */s o s ə l/*

2.2.4. Missing and Confused Approximants

Articulation of the English approximant */r/* involves lip rounding and retroflexion. */r/* is absent from Cantonese and is often substituted with */w/* (rounded approximant) or */l/* (lateral approximant), e.g.:

- "rate" */ret* / vs. "wait" */wet/*
- "very" */v ɛ r i* / as */v ɛ l i/* or */w ɛ l i/*

2.2.5 Confusion among /n/ and /l/

In colloquial Cantonese, it is often acceptable to substitute /*n*/ with /*l*/, e.g. e.g.你(you) */nei/* pronounced as 理 (logic) */lei/*. Hence, Cantonese learners often perform such substitutions in English, e.g. "nine" */n ay n/* as "line" */l ay n/*. This is perceived as a prominent mispronunciation.

	Bilabial	Labio-dental	Labio-velar	Dental	Alveolar	Pre-Palatal	Palatal	Velar	Glottal
Plosive	p , p^h				t, t^h			k, k^h,k^w, k^{wh}	
Affricate					ts, ts^h				
Nasal	m				n			ŋ	
Fricative		f			s				h
Approximant	w						j	(w)	
Lateral Approximant					l				

Table 1. Consonants in Cantonese, organized according to the manner and place of articulation [9].

	Bilabial	Labio-dental	Labio-velar	Dental	Alveolar	Pre-Palatal	Palatal	Velar	Glottal
Plosive	p , p^h, b				t, t^h, d			k, k^h, g	
Affricate						tʃ, dʒ			
Nasal	m				n			ˉ	
Fricative		f, v		θ,ð	s, z	ʃ, ʒ			h
Approximant	w				r		j	w	
Lateral Approximant					l				

Table 2. Consonants in American English [10].

2.2.6 Phonotactic Constraints

The Cantonese syllable has a simple [C]V[C] structure, where the optional syllable onset contains one consonant (except for /kw/ or /k^hw/) and the optional syllable coda also contains one consonant. The English syllable also has optional onset and coda, but each may contain up to three consonants, such as in the word "strengths" /strɛŋθs/. Cantonese learners frequently mispronounce English consonant clusters, either with *vowel insertion* in the word-final position, e.g. "kissed" */kɪst/* becomes */kɪstɨ/;* or *consonant deletion*, e.g. "professor" */prəfɛsə/* becomes */pou-fɛ-sa/.*

3. DATA COLLECTION

Corpus collection is important for diagnostic analysis of mispronunciations in CALL systems [3, 14]. We adopt the Sennheiser PC155 headset which consists of a noise-canceling uni-directional microphone and built-in sound card. Recordings are conducted from three sites: a sound-proof recording room [1] and two study rooms [2] (without sound-proofing). We selected speakers based on the criteria that their mother tongue is the Cantonese dialect, they have learned English for at least 10 years and their English pronunciation are deemed intermediate to good by the English teachers in our university. Each speaker is asked to verify the recording quality of their utterances through playback as well as waveform visualization (e.g. to ensure no clipping). Thus far we have recordings from 21 speakers (9 male, 12 female) and the recording effort will continue and the number of speakers will increase. The data was digitized at 16-bit per sample and a sampling rate of 16 kHz. The average SNR of the recordings are 37.6dB for the sound-proof room and 36.7dB for the study rooms.

The recording text prompts include several types:

(i) *The North Wind and the Sun*, a classic example of Aesop's Fables;

(ii) Minimal pairs, confusable word groups and phonemic sounds that are designed by English teachers in our university; and

(iii) Sentences from the TIMIT database [11] to cover a variety of phonetic contexts.

The current work is a *feasibility study* of the proposed methodology for deriving salient learner's mispronunciations. Hence we focus our subsequent experimentation and analyses on the recordings of "*The North Wind and the Sun*," because this piece is commonly used by linguists to exemplify languages. There are a total of 113 words in this story and the lexicon size is 64 words. The results of our study should be generalizable to other speech recordings, such as the two remaining ones mentioned above.

4. SPEECH RECOGNITION EXPERIMENTS

We develop a speech recognizer with cross-word triphone HMMs that contain 2000 states with 12 Gaussian mixtures per state. This implementation is based on the HTK toolkit [15]. These are trained on the TIMIT training set [12] which contains a total of 4620 sentences recorded by 462

[1] In the Human-Computer Communications Laboratory, CUHK.

[2] In the CUHK Independent Learning Center [13].

speakers from eight dialect regions of the US.

The test set consists of speech recordings made by the 21 Cantonese learners. Each recording of "The North Wind and the Sun" is segmented into six utterance files, each corresponding to a single sentence. Each utterance consists of 10 to 26 words in natural, continuous speech.

We adopt the TIMIT dictionary[3] and extended it to include *predicted word mispronunciations*, by incorporating the possible phone-to-phone mappings based on the phonetic transfer effects as described in Section 2. For example, the confusion of "there" /*dh eh r*/ versus "dare"/*d eh r*/ (as mentioned in Section 2.2.3) signifies the possible mapping between /*dh*/ ←→ /*d*/. Each confusable phone is mapped to zero (deletion), one or more phones (substitutions). Hence the 64 words in the story were mapped to over 300 pronunciations in all, where the number of pronunciations for a given word may range from 1 to 30. For example, the word "cloak" has */k l ow k/* as its correct pronunciations, as well as */k ow k/, /k l ao k/,* etc. as potential erroneous variants. The extended pronunciation dictionary is used to generate a word network as an input to the speech recognizer for decoding the testing utterances.

Unlike conventional automatic speech recognition, our current task decodes speech for which the reading text is known [6]. The recognizer decodes the best phone sequence given the known sequence of words and their possible pronunciations from the extended dictionary. Hence, ASR can be used to pinpoint erroneous word pronunciations that included in the extended pronunciation dictionary.

5. DERIVING SALIENT MISPRONUNCIATIONS

We tabulated the statistics of phonetic confusions by comparing the reference phonetic transcriptions with the best-scoring phonetic sequence in the recognizer's hypothesis. Among the 64 words in the story, we found that 7 of them are pronounced correctly by all 21 speakers. Of the remaining words, we extract mispronunciations that occur with a relative frequency of at least 25% per word. This procedure extracts prominent mispronunciations for 31 words in the lexicon. We organize these mispronunciations according to the language transfer effects as described in Section 2:

(i) Salient vowel confusions, e.g.
"cloak*"* */k l ow k/* → */k l ao k/*
"last" */l ae s t/* → */l ah s t/*
"traveler" */t r ae v el axr/* → */t r aa f el axr/*
"disputing"*/d ix s p y uw t ix ng/* → */d ix s p uw t ix ng/*
These may lead to semantic confusions between "cloak" and "clock", or "last" and "lust", etc.

(ii) Confusions with long and short vowels, e.g.
"could" */k uh d/* → */k uw d/*
"him" */hh ih m/* → */hh iy m/*
"wind" */w ih n d/* → */w iy n d/*
These may lead to semantic confusions between "could" and "cooed", "him" and "Heem", or "wind" and "weaned".

(iii) Confusions with voiced plosives, e.g.
"agreed" */ax g r iy d/* becomes → */ax g r iy t/*
"hard" */hh aa r d/* → */hh aa t/* or */hh aa r t/*
These may lead to semantic confusions among "hard", "hot" and "heart".

(iv) Confusions with labiodental fricatives, e.g.
"traveler" */t r ae v el axr/* → */t r aa f el ax/*
"gave" */g ey v/* → */g ey f/*

(v) Confusions with dental fricatives, e.g.
"north" */n ao r th/* → */n ow l f/*
"than" */dh ae n/* → */d ae n/*
"that" */dh ae t/* → */d ae t/*
"the" */dh ax/* → */d ax/*
These may lead to semantic confusions between "than" and "Dan", or "that" and "Dad".

(vi) Confusions with alveolar voiced fricatives, e.g.
"as" */ae z/* → */ae s/*
"his" */hh ih z/* → */hh ih s/*
"was" */w ax z/* → */w ax s/*
"closely" */k l ow z l iy/* → */k l ow s l iy/*
These may lead to semantic confusions between "his" and "hiss", or "was" and "what's".

(vii) Confusions with the approximant-alveolar /r/, e.g.
"hard" */hh aa r d/* → */hh aa t/*
"more" */m ao r/* → */m ao/*
"north" */n ao r th/* → */n ow l f/* or */n ao f/*
"stronger" */s t r ao ng axr/* → */s t r ao ng ax/*
"traveler" */t r ae v el axr/* → */t r aa f el ax/*
"warm" */w ao r m/* → */w ao m/*
"warmly" */w ao r m l iy/* → */w ao m l iy/*

(viii) Confusions due to phonotactic constraints – vowel insertions, e.g.
"agreed" */ax g r iy d/* → */ax g r iy d ax/*
"wrapped" */r ae p t/* → */r ae p t ax/*
These may lead to semantic confusions between "wrapped" and "raptor".

(ix) Confusions due to phonotactic constraints – consonant deletions, e.g.
"attempt" */ax t eh m p t/* → */ax t eh m/ or /ax t eh m p/*
"blew" */b l uw/* → */b uw/*
"first" */f er s t/* → */f er s/ or /f er t/*
"should'' */sh uh d/* → */sh uh/*
"out" */aw t/* → */aw/*
"wind" */w ih n d/* → */w iy n/*
"succeeded" */s uh k s iy d ix d/* → */s uh s iy d ix t/*
These may lead to semantic confusions between

[3] Henceforth we will use Darpabet instead of IPA.

"blew" and "boo", "first" and "furs", or "wind" and "wean".

We believe the above mispronunciations are very useful for the design of remedial instructions for pronunciation improvement.

6. CONCLUSIONS AND FUTURE WORK

This paper describes an initial effort to derive salient pronunciation errors made by Cantonese learners of English. The proposed methodology is grounded on the theory of language transfer and involves phonological comparisons between Hong Kong Cantonese (L1) and American English (L2) across the phonetic and phonotactic levels. We identify major disparities across the two languages, which are believed to heighten the perceived phonological interference of transfer features and cause mispronunciations. Systematic phonological comparisons enable us to predict possible phonetic confusions that lead to word mispronunciations in second language acquisition.

Our methodology then incorporates the predicted phonetic confusions to extend a pronunciation lexicon with possible word mispronunciations. The extended pronunciation lexicon is used to produce a confusion network in recognition of the English speech recordings of Cantonese learners. We refer to the statistics of the erroneous recognition outputs to derive salient mispronunciations that stipulates the predictions based on phonological comparison.

Initial experimentation based on the speech recordings from 21 Cantonese learners of English shows that the proposed methodology is effective in deriving salient mispronunciations. These will be useful in the design of remedial instructions in language learning. In the near future, we plan to apply the proposed methodology to an extended speech corpus, as well as develop mispronunciation detection schemes based on the trained recognizer. As we adapt this speech recognizer for mispronunciation detection, we will also measure the detection performance under real use. We will also generalize the proposed methodology to derive mispronunciations made by Mandarin (L1) learners of American English (L2).

7. ACKNOWLEDGMENTS

This work is partially supported the CUHK Teaching Development Grant. The authors thank Professor Eric Zee and Dr. Wai-Sum Lee of the City University of Hong Kong, as well as Professor Yoshinori Sagisaka of Waseda University, for many helpful discussions. The authors also wish to thank Ms. Pauline Lee of CUHK's Independent Learning Center and Dr. Peter Clark of CUHK's English Language Teaching Unit for their support.

8. REFERENCES

[1] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "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] Ordinate Corporation, "The PhonePass Test", Menlo Park, CA, January 1998.

[3] W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein, "Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English", Conference on Speech Technology in Language Learning, Marholmen, Sweden, 1998.

[4] S. M. Witt and 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.

[5] B. Mak, M. H. Siu, M. Ng, Y. C. Tam, Y. C. Chan, K. W. Chan, K. Y. Leung, S. Ho, F. H. Chong, J. Wong, J. Lo "PLASER: Pronunciation Learning via Automatic Speech Recognition", Proceedings of HLT-NAACL, 2003.

[6] J. Mostow, "Is ASR accurate enough for automated reading tutors, and how can we tell?", International Conference on Spoken Language Processing ICSLP, 2006.

[7] Y. Tsubota, T. Kawahara, and M.Dantsuji, "Recognition and verification of English by Japanese students for computer-assisted language learning system", Proceedings of ICSLP, pp.1205 – 1280, 2002.

[8] Zue, V., Notes on Speech Spectrogram Reading, MIT Summer Course, 1991.

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

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

[11] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, "An Acoustic-Phonetic Data Base", J. Acoust. Soc. Am. 81, Suppl 1, 1987.

[12] J. D. O'Connor "Better English Pronunciation", Cambridge University Press, 1980.

[13] Independent Learning Center, CUHK, http://www.ilc.cuhk.edu.hk/english/index.htm.

[14] E. Atwel, P. Howarth and C. Souter, "The ISLE Corpus: Italian and German Spoken Learners", Computers in English Linguistics, ICAME Journal, 2003.

[15] S. J. Young, J. Odell, D. Ollason and P.C. Woodland, "The HTK book Entropic", Cambridge Research Laboratory, 1996.

Appendix A.

Cantonese vowels and diphthongs in the context of syllable pronunciations of Chinese characters. Meanings of the characters are in parentheses.

Four short vowels			Seven long vowels		
ɪ	/sɪk/	色 (color)	i	/si/	絲 (silk)
ɐ	/sɐp/	濕 (wet)	y	/sy/	書 (book)
ʊ	/sʊk/	叔 (uncle)	ɛ	/sɛ/	借 (borrow)
θ	/sθt/	恤 (shirt)	œ	/hœ/	靴 (boot)
			a	/sa/	沙 (sand)
			ɔ	/sɔ/	梳 (comb)
			u	/fu/	夫 (husband)
Ten diphthongs					
ai	/gai/	佳 (good)	θy	/sθy/	稅 (tax)
ɐi	/sɐi/	西 (west)	ɔi	/hɔi/	開 (open)
au	/gau/	交 (give)	ui	/fui/	灰 (gray)
ɐu	/sɐu/	收 (receive)	iu	/siu/	燒 (burn)
ei	/sei/	四 (four)	ou	/dou/	刀 (knife)

Appendix B.

American English vowels and diphthongs in the context of an English word.

i	beat	o	boat
ɪ	bit	ʊ	book
e	bait	u	boot
ɛ	bet	ɚ	Burt
æ	bat	aɪ	bite
a	Bob	ɔɪ	Boyd
ɔ	bought	aʊ	bout
ʌ	but	ə	about