Development of a Cantonese-English Code-mixing Speech Corpus

Joyce Y. C. Chan, P. C. Ching and Tan Lee

Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong SAR, China
{ycchan, pcching, tanlee}@ee.cuhk.edu.hk

Abstract

This paper describes the design and compilation of the CUMIX Cantonese-English code-mixing speech corpus. Code-mixing is a common phenomenon in many bilingual societies and it usually involves at least two different languages within one utterance. In Hong Kong, people usually mix English words and phrases with Cantonese in their daily conversation. Although there are many monolingual corpora of Cantonese and English, code-mixing speech database of these two languages is not available. The aim of developing this corpus is to study of the effect of Cantonese accents in English, the design of effective language boundary detection algorithm in code-mixing utterances [1], and evaluation of the performance of code-mixing speech recognizers.

1. Introduction

According to John Gumperz[2], the definition of code-switching is "the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-system". In Hong Kong, code-switching tends to be intra-sentential and switching involving linguistic units above the clause level is rare, hence the preference for the term "code-mixing" in many studies [3]. Although there is admittedly a grey area between (inter-sentential) code-switching and (intra-sentential) code-mixing, "code-mixing" is a more preferable term to describe the typical language behaviour of the average Hong Kong bilinguals [4].

Hong Kong is a truly international city and most people are Cantonese-English bilinguals. Code-mixing between Cantonese and English is a common practice of them, for both speech as well as written text. Cantonese is the mother tongue of most residents in Hong Kong. It is usually the matrix language while English is the embedded language that is often used to better describe meanings, feelings and phenomena. However, the English words uttered by many local people do contain Cantonese accent, which makes automatic speech recognition difficult.

In order to study the effect of Cantonese accent on the English words, as well as to provide data for training the English phone models, monolingual English or Cantonese-English code-mixing data is required. Large amount of code-mixing speech data, rather than monolingual English, is preferred since people may pronounce the same word totally different in the two situations. To evaluate the code-mixing speech recognizer, code-mixing speech data is necessary and monolingual Cantonese data are also required as the baseline for performance comparison.

2. Phonological structure of Cantonese and English

The phonological structures of Cantonese and English are quite different because they come from two different language families. Cantonese is one of the major Chinese dialects, which is a Sino-Tibetan language. It is monosyllabic in nature and has a general syllable structure C1VC2, where C1 and C2 are optional consonants and V is either simple vowel or diphthong. All the Cantonese syllables are of the canonical forms V, CV, CVC or VC [5]. On the other hand, English is Indo-European language and the phonological structure is much more complicated than Cantonese. In English discourse, over 80% of the syllables are of the canonical form of Cantonese, and the remainings are C, CC, CCV, VCC, CCCV, CCCVCC, etc [6].

2.1. Cantonese accent in the embedded English words

In code-mixing utterances, the words in embedded language may be pronounced with heavy accent of the matrix language. This phenomenon is called borrowing [7]. For Cantonese speakers, the borrowing words are pronounced with the following characteristics [8]:

- Softening or dropping the second consonant in a CC sequence, e.g. plan /p l ae n/ is pronounced as /p ae n/
- Softening or dropping the final stop consonant e.g. check /ch eh k/ is pronounced as /ch eh/
- Adapting a monosyllabic word with fricative endings to produce a disyllabic, e.g. notes /n ow t s/ is pronounced as /n ow t s iy/
- Retroflex such as /r/ is read as /l/ sound or /w/ sound, e.g. pressure /p r eh sh er/ is pronounced as /p l eh sh er/, and repeat /r iy p iy t/ is pronounced as /w iy p iy t/
- If the phone exists in English only but not in Cantonese, they will be pronounced as the similar phones in Cantonese, such that /th/ becomes /f/, and /eh/ becomes /ae/

2.2. Phone change and syllable fusion in Cantonese

Although Cantonese is the mother tongue of most people in Hong Kong, not all of them are able to pronounce all the Cantonese words correctly. One of the reasons is that Hong Kong people do not use romanization systems when they learn Chinese or Cantonese. People may not know the correct pronunciation of the words, and confuse a phoneme with the other. For example, the word "我" /ngo3/ are

sometimes pronounced as /o3/, i.e. the initial /ng/ is dropped. The commonly confused Cantonese phonemes are listed in the following table:

Table 1: Commonly confused phonemes in Cantonese

The original phoneme	Realize as another phoneme
Initial /ng/ (e.g. ngo5, 我)	Null initial (e.g. o3)
Initial /n/ (e.g. naam4, 男)	Initial /l/ (e.g. laam5, 藍)
Final /ng/ (e.g. hong4, 航)	Final /n/ (e.g. hon4, 寒)
Final /k/ (e.g. bak3, 百)	Final /t/ (e.g. bat3, 八)

Besides, syllable fusion may occur in fast speech. The pronunciation of the second syllable of disyllabic words may be ignored or changed. For example, the word "知道" /zi1 dou3/ may be pronounced as /zi1 ou3/, "今日" /gam1 jat6/ becomes /gam1 mat6/ [9].

3. Corpus design and data collection

The purpose of this corpus is to provide speech data for the training of Cantonese accented English acoustic models, and evaluate the performance of the Cantonese-English code-mixing speech recognition system. The data can also be used to study on automatic language identification within code-mixing utterances [1].

3.1. Training data and testing data

3.1.1. Training data

The training data includes four types of speech data: 1) Cantonese-English code-mixing utterances, 2) monolingual English words, 3) English numbers and 4) English alphabets. It includes speech data from 20 male and 20 female speakers. Each speaker read 200 code-mixing utterances, 100 English words or phrases, as well as English numbers and English alphabets. Each code-mixing utterance includes one English segment only.

The monolingual English words are those commonly used in code-mixing. Large proportion of the speech data are Cantonese-English code-mixing, so that the pronunciation of the English words include more Cantonese accents. It is because when people are code-mixing, they usually adapt the code-switch words to their mother tongue. However, if they are speaking in monolingual English, they will try to speak every word clearly, so the pronunciation of the words will be different from the code-mixing one. For example, when people speak the word "notes", if it is in an English sentence, it will be pronounced as /n ow t s/. However, if it is in code-mixing, they may adapt it to disyllabic, such that the pronunciation become /n ow t s iy/. This type of adaptation occurs in code-mixing only but not monolingual English, even for the same speaker.

There are about 8000 code-mixing utterances in total, which is inadequate for training good Cantonese acoustic models, hence other corpus is needed. At CUHK, we have developed the CUSENT Cantonese speech corpus, which mainly contains read newspaper content. It is phonetically rich under various contexts, and the recording environment is the same as this corpus [5]. Although the lexicons in written Cantonese are different from those in spoken Cantonese, the syllables involved are similar. The data in CUMIX can be applied for adapting the written Cantonese to spoken Cantonese.

3.1.2. Testing data

The testing data is for performance evaluation of the code-mixing speech recognizer or language identification system. Part of the data can be used for development purpose, such as acoustic model adaptation and parameters tuning. Each speaker read 120 code-mixing utterances, and about 90 monolingual Cantonese utterances. Among the 120 code-mixing utterances, 110 of them include single English segment only, and the remaining 10 utterances include two English segments. The monolingual Cantonese utterances are identical to the code-mixing utterances except that the code-switch words are replaced by their Cantonese equivalent. If Cantonese equivalent does not exist, the monolingual version of the utterance will not be recorded, thus the number of monolingual Cantonese utterance is less then those in code-mixing. The following is an example of the code-mixing utterance and monolingual Cantonese utterance, where the code-switch word "bonus" is replaced by its Cantonese equivalent "花紅" in the monolingual case:

Table 2: Example of code-mixing utterance and monolingual Cantonese utterance

Code-mixing utterance									
ngo5	gok3-dak1	gam1-nin4	jau5	B OW N AH S		ge3	gei1-wui6	hou2	miu5-mong4
我	覺得	今年	有	bonus		嘅	機會	好	渺茫。
Monolingual Cantonese utterance									
ngo5	gok3-dak1	gam1-nin4	jau5	faa1-hung4		ge3	gei1-wui6	hou2	miu5-mong4
我	覺得	今年	有	花紅		嘅	機會	好	渺茫。
English Translation									
I	believe	this year	have	bonus		DET	chance	very	low
(I believe that there is low probability to have bonus this year)									

3.2. Code-mixing text material

Appropriate text materials are needed to reflect and cover the most prominent code-mixing scenario. The conventional way of obtaining these materials is to extract them directly from generally accessible text database such as local newspapers and books. However, this is not the case for Cantonese. Cantonese is a dialect and spoken Cantonese is noticeably different from written Cantonese. The grammar is similar but the lexicon selection is quite different. Here is an example:

Written Cantonese:	我	明天	不用	上學。
Spoken Cantonese:	我	聽日	唔駛	返學。
English Translation:	I	tomorrow	need not	go to school

(I need not go to school tomorrow)

The text materials for recording are in spoken Cantonese, which is a 'low' language, such that most of the local newspapers and books do not use it. So as to collect adequate text materials which involve code-mixing of spoken Cantonese and English, other sources such as newsgroup and online diary are also included. Utterances being used in the previous researches related to Cantonese-English code-mixing / code-switching are also considered [10].

The followings are criteria of selecting the code-mixing text material:

- frequency of occurrence of the code-switched words
- part-of-speech (POS) of the code-switch words
- the 'length' of the code-switch words, e.g. single alphabet, abbreviations, single word, multiple words, phrase

POS distributions and word length distributions are considered in the speech corpus. There are 1047 unique code-switch words in the training data, and 1069 in the testing data. Each code-switch word appears 4 to 12 times in the training data, which depends on frequency of occurrences in daily conversation. There are in total 2087 unique code-mixing utterances in the training data, and 2256 in the testing data. The following table lists the POS distributions and word length distributions:

Table 3: POS distribution of the code-switch words

Parts-of-speech	No. of utterances	Proportion
Noun / Noun Phrase	7723	62.3%
Verb / Verb Phrase	2778	22.4%
Adjectives / Adverbs	1771	14.3%
Phrases / Others	128	1.0%

Table 4: Word length distribution of the code-switch words

Word length	No. of utterances	Proportion
1	9295	74.96%
2	2197	17.72%
3	94	0.76%
4	16	0.13%
abbreviation	692	5.58%
ENG-CAN-ENG	106	0.85%

Some of the code-switch words are abbreviations or the "short form" of the English words. For example, the word "introduction" /ih n t r ah d ah k sh ah n/ is usually modified to disyllabic and become "intro" /ih n t r ah/ in Cantonese-English code-mixing utterances.

Besides, there are also English-Cantonese-English patterns in Cantonese-English code-mixing utterances. The Cantonese words involved are usually single character word, such as "唔" /ng4/ (not). For example, people will say "un 唔 understand" (understand or not?), or "mind 唔 mind"(mind or not?) [10].

3.3. Data collection

3.3.1. Recording environment

The speech data are collected in a closed silent recording room. The recordings are expected to be good quality clean speech data. The speakers are requested to read the utterances naturally and fluently, and use their normal pronunciation for the English words. Therefore, borrowing may occur in the English words, and phone change or syllable fusion may occur in the Cantonese words. The speakers have to check if they know the pronunciation of all the code-switch words. Correct pronunciation will be told to ensure the quality of the code-switch words. The speakers were left alone in the recording room to do the recording themselves without assistance or intervention. They read the utterance prompted on a computer screen. The recording of each utterance is terminated automatically with silence detection. The speakers were requested to record the utterances again if the noise level is too high or the pronunciation of the code-switch words is too far away from the correct one.

The recording is collected using a high quality, unidirectional dynamic head-worn microphone. The microphone position is about 1 inch from the corner of the mouth, so as to eliminate explosive breath sounds. The signal passes through a pre-amplification mixer and sampled by a DAT recording at 48kHz and 16bit. The sampled digital data is then down sampled to 16kHz, and transferred through SCSI interface to computer and stored as disk files. The data collection set-up is shown in Figure 1.

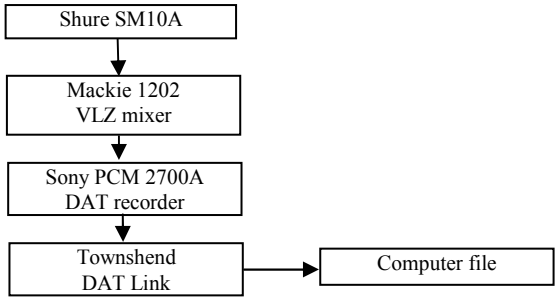

Figure 1: Block diagram for data collection set-up

3.3.2. Background of speakers

This speech corpus includes speech data from 80 speakers, 40 male and 40 female. 20 male and 20 female speakers are for training data, and the remaining are for testing data. All the speakers are native Cantonese speakers and able to speak fluent English. They are undergrad students or graduate students from the Chinese University of Hong Kong, aged between 19 and 26.

4. Verification and annotation

Verification is performed in two stages. Stage one is done by generally trained assistants to mark out all incorrectly spoken and noisy data. The speakers are requested to record these data again.

Stage two of the verification process is performed by experts in phonetics. Pronunciation of the code-switch words is labeled and the Cantonese transcriptions will be corrected. If deemed uncorrectable, the corresponding data will be discarded.

Language boundaries (in ms) are also labeled manually for language identification. It uses the MLF format of HTK, includes the time alignment for silence, Chinese content and English content. For data annotation, orthogorphic transcriptions in BIG5 code and phonemic transcriptions are provided. Phonemic transcriptions of English words are based on the CMU Pronouncing Dictionary version 0.6. ARPABET is used and the phoneme set has 39 phonemes, lexical stress is ignored. Cantonese phonemes are labeled with LSHK. Spoken Cantonese includes Hong Kong-specific characters which are not contained in the BIG5 standard

character set, thus the Hong Kong Supplementary Character Set (HKSCS) must be installed in order to view the text.

5. Data organization

In order to facilitate the research and development of Cantonese-English code-mixing speech processing, the corpus data will be distributed in electronic form through CDROM. All speech data will be accompanied by both orthogorphic and phonemic transcriptions, as well as language boundary information. The following table will summarize the format and of the corpus:

Table 5: *Format of the CUMIX corpus*

Speech data format	NIST SPHERE
Sampling rate	16kHz
Precision	16 bit per sample
Orthographic transcription	BIG5 code (HKSCS)
Phonemic transcription	Cantonese: LSHK English: ARPABET
No. of speakers	Training: 20M, 20F Testing: 20M, 20F
No. of code-switching / monolingual Cantonese utterances	Training: 8000 Testing: 4800 / 3600
Corpus size (include silence time)	Training: 9 hr Testing: 8 hr

The data layout structure is simple. Each speaker will be given a speaker code which includes the speaker ID. All data obtained from the same speaker will be placed under the same directory which is named according to the speaker code. The speaker ID starts with either 'M' or 'F' for speaker gender, and then followed by a two-digit decimal integer. For example, data for a male speaker may be put under the directory M03. Basic information of the speaker, such as age, gender, and major subject in university are also included in the directory.

6. Discussion

Without the CUMIX training data, we have to use two monolingual speech corpora to train the acoustic models, which lead to the degradation in code-mixing speech recognition accuracy. Speech recognizer with acoustic models trained by the Cantonese speech corpus CUSENT and the American English speech corpus TIMIT [11], have word accuracy (syllable accuracy for Cantonese) 51.31% for code-mixing testing data in the CUMIX. However, if acoustic models with the same features (12 MFCC, normalized energy, first and second derivatives of the parameters) are trained by CUSENT (20,000 utterances) and CUMIX (8,000 utterances, 4,000 English words), the word accuracy becomes 55.67%, with 4% absolute improvement. The CUMIX training data provides spoken Cantonese and Cantonese accented English speech for training the acoustic models as well as phone change analysis. It also provides adequate data for tuning parameters for language boundaries detection algorithms in code-mixing utterances. [1]

7. Conclusion

In summary, the collection of a Cantonese-English code-mixing database for speech processing has been launched. Small amount of Cantonese accented English data has been collected for training the English models and analysis on the phone change in these English words. Other phonetically rich continuous Cantonese speech corpus such as CUSENT should be used to train the Cantonese models for LVCSR applications, and the spoken Cantonese data in this corpus provide data for adapting the written Cantonese in CUSENT to spoken Cantonese. Cantonese-English code-mixing data can evaluate the performance of the code-mixing speech recognizer, and the relevant monolingual Cantonese data can be compared as a baseline result.

With these speech data, more in depth investigation on code-mixing can be conducted. After validation of the transcriptions and speech data, the corpus will be public available. More Cantonese accented English data may be required and spontaneous code-mixing speech should be collected for further research on it.

8. References

[1] Joyce Y.C. Chan, P. C. Ching, Tan Lee, Helen M. Meng, "Detection of Language Boundary in Code-Switching Utterances by Bi-phone Probabilities", *Proc. of ISCSLP 2004,* 293-296, Hong Kong, 2004

[2] John Gumperz, *Discourse Strategies*, p.59, Cambridge University Press, 1982

[3] David C. S. Li, "Cantonese-English code-switching research in Hong Kong: a Y2K review", *World Englishes, Vol. 19, No.3, p.305-322*, Blackwell Publishers Ltd., 2000

[4] A. Tse, "Some observations on code-switching between Cantonese and English in Hong Kong", *Working Papers in Languages and Linguistics, Vol. 4, p.101-108*, Department of Chinese, Translation and Linguistics, City Polytechnic of Hong Kong, 1992

[5] P.C. Ching, K.F. Chow, Tan Lee, Alfred Y.P. Ng and L.W. Chan, "Development of a large vocabulary speech database for Cantonese", *Proc. of ICASSP 1997*, Munich, Germany, 1997

[6] Mirjam Wester, "Syllable Classification using Articulatory-Acoustic Features", *Proc. of Eurospeech 2003*, pp. 233-236, Geneva, Switzerland, 2003

[7] Mimi Chan, Helen Kwok, *A study of lexical borrowing from English in Hong Kong Chinese*, Centre of Asian Studies, University of Hong Kong, Hong Kong, 1990

[8] Ping Li, "Spoken Word Recognition of Code-Switched Words by Chinese-English Bilinguals", *Journal of Memory and Language*, Vol. 35, p.757-774, 1996

[9] Peggy W. Y. Wong, "Syllable Fusion and Speech Rate in Hong Kong Cantonese", *Proc. of Speech Prosody 2004, p.* 255-258, Nara, Japan, 2004

[10] Brian Hok-Shing Chan, *Code-mixing in Hong Kong Cantonese-English bilinguals : constraints and processes*, M. A. Thesis, The Chinese University of Hong Kong, 1992.

[11] Victor W. Zue, Stephanie Seneff, "Transcription and Alignment of the TIMIT Database", *The 2nd Symposium on Advanced Man-Machine Interface through Spoken Language*, Hawaii, 1988