Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 281-304

Automatic Recognition of Cantonese-English Code-Mixing Speech

Joyce Y. C. Chan[*], Houwei Cao[*], P. C. Ching[*], and Tan Lee[*]

Abstract

Code-mixing is a common phenomenon in bilingual societies. It refers to the intra-sentential switching of two different languages in a spoken utterance. This paper presents the first study on automatic recognition of Cantonese-English code-mixing speech, which is common in Hong Kong. This study starts with the design and compilation of code-mixing speech and text corpora. The problems of acoustic modeling, language modeling, and language boundary detection are investigated. Subsequently, a large-vocabulary code-mixing speech recognition system is developed based on a two-pass decoding algorithm. For acoustic modeling, it is shown that cross-lingual acoustic models are more appropriate than language-dependent models. The language models being used are character tri-grams, in which the embedded English words are grouped into a small number of classes. Language boundary detection is done either by exploiting the phonological and lexical differences between the two languages or is done based on the result of cross-lingual speech recognition. The language boundary information is used to re-score the hypothesized syllables or words in the decoding process. The proposed code-mixing speech recognition system attains the accuracies of 56.4% and 53.0% for the Cantonese syllables and English words in code-mixing utterances.

Keywords: Automatic Speech Recognition, Code-mixing, Acoustic Modeling, Language Modeling

1. Introduction

Code-switching and code-mixing are common phenomena in bilingual societies. According to John Gumperz (Gumperz, 1982), the definition of code-switching is “the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical

[*] Department of Electronic Engineering, The Chinese University of Hong Kong
E-mail: {ycchan, hwcao, pcching, tanlee}@ee.cuhk.edu.hk

systems or sub-systems". Different combinations of languages are found in code-switching, for examples, Spanish-English in United States, German-Italian and French-Italian in Switzerland, and Hebrew-English in Israel (Auer, 1998). In Taiwan, code-switching between Chinese dialects, namely Mandarin and Taiwanese, has become common in recent years (Chen, 2004). Hong Kong is an international city where many people, especially the younger generation, are Cantonese and English bilinguals. English words are frequently embedded into spoken Cantonese. The switching of language tends to be intra-sentential, and it rarely involves linguistic units above the clause level. Hence, the term code-mixing is usually preferred (Li, 2000). In this case, Cantonese is the primary language, also known as the matrix language, and English is the secondary language, usually referred to as the embedded language (Halmari, 1997).

Automatic speech recognition (ASR) is one of the key technologies in spoken language processing. An ASR system converts an input speech waveform into a sequence of words. Recently, ASR for multilingual applications has attracted great interest (Schultz & Kirchhoff, 2006). In state-of-the-art ASR systems, the input speech is assumed to contain only one language and the language identity is given. These systems are not able to handle code-mixing speech, which differs significantly from monolingual speech spoken by native speakers. This calls for special consideration in the design of acoustic models, lexical and language models, and in the decoding algorithm.

There have been two different approaches to code-switching or code-mixing speech recognition (Lyu *et al.,* 2006; Chan *et al.,* 2006). The first approach involves a language boundary detection (LBD) algorithm that divides the input utterance into language-homogeneous segments. The language identity of each segment is determined, and the respective monolingual speech recognizer is applied. LBD for mixed-language utterances was studied by Wu *et al.* (2006) and Chan *et al.* (2004). Language-specific phonological and acoustic properties were used as the primary cues to identify the languages. The second approach aims to develop a cross-lingual speech recognition system, which can handle multiple languages in a single utterance. The acoustic models, language models, and pronunciation dictionary are designed to be multi-lingual and cover all languages concerned. In Lyu *et al.* (2006), automatic recognition of Mandarin-Taiwanese code-switching speech was investigated. It was found that Mandarin and Taiwanese, both of which are Chinese dialects, share a large percentage of lexicon items. Their grammar was also assumed to be similar. A one-pass recognition algorithm was developed using a character-based search net. It was shown that the one-pass approach outperforms LBD-based multi-pass approaches. In You *et al.* (2004), a mixed-lingual keyword spotting system was developed for auto-attendant applications. The keywords to be detected could be in either English or Chinese.

This paper presents a study on automatic speech recognition of Cantonese-English code-mixing speech. Part of the work was reported in Chan *et al.* (2006). Our study covers all components of an ASR system, including acoustic models, language models, pronunciation dictionary, and search algorithm. Different approaches to LBD are also investigated. By understanding the linguistic properties of monolingual Cantonese and English, as well as code-mixing speech, the major difficulties in code-mixing speech recognition are revealed and possible solutions are suggested. We propose a two-pass recognition system, in which the acoustic and linguistic knowledge sources are integrated with language boundary information. Simulation experiments are carried out to evaluate the performance of the whole system as well as individual system components.

2. Difficulties in Code-mixing Speech Recognition

2.1 Linguistic Properties of Cantonese and English

Cantonese is a Chinese dialect. It is spoken by tens of millions of people in the provinces of Guangdong, Guangxi, Hong Kong, and Macau. A Chinese word in its written form is composed of a sequence of characters. In Cantonese, each Chinese character is pronounced as a monosyllable carrying a specific lexical tone (Ching *et al.,* 2006). English is one of the most popular languages in the world. An English word is written as a sequence of letters. In spoken form, each word may consist of several syllables, some of which are designated to be stressed. Table 1 shows a pair of example words in Cantonese and English.

Table 1. Examples of Cantonese and English words in written and spoken format.

Written (orthographic transcription)	Spoken (phonetic transcription)
產生	/ts^h a n/ /s ɐ ŋ/
produce	/p r ə ˈd j uː s/

Syllables can be divided into smaller units, namely consonants (C) and vowels (V). Cantonese syllables take the structures of V, CV, CVC, or VC (Ching *et al.,* 1994). If tonal difference is not considered, the number of distinct Cantonese syllables is around 600 (Ching *et al.,* 2006). The syllable structure in English is more complicated than that in Cantonese. Although many English syllables share the same canonical forms as given above, there also exist combinations like CCV, VCC, CCCV, and CCCVCC (Wester, 2003), which are not found in Cantonese.

There are 22 consonants and 22 vowels (including diphthongs) in Cantonese, and 24 consonants and 14 vowels in American English (Ching *et al.,* 1994; Ladefoged, 1999). Table 2 lists the IPA (International Phonetic Alphabet) symbols of these phonemes. Some of the phonemes in the two languages are labeled with the same IPA symbols by phoneticians, meaning that they are phonetically very close. Some of the other phonemes are also considered to be very similar although they are labeled differently in the two languages, *e.g.*, /**au**/ in Cantonese and /**aʊ**/ in English.

Table 2. Phonemes of Cantonese and English. The phonemes that are labeled with the same IPA symbols in both Cantonese and English are listed first and boldfaced.

Cantonese phonemes	
IPA symbol	Example
p	[p a] （爸）
m	[m a] （媽）
f	[f a] （花）
t	[t a] （打）
tʃ	[tʃ y] （朱）
n	[n a] （拿）
s	[s a] （沙）
ʃ	[ʃ y] （書）
l	[l a] （啦）
j	[j ɐu] （憂）
k	[k a] （加）
ŋ	[p^h a ŋ] （烹）
w	[w a] 蛙
h	[h a] （蝦）
ɪ	[s ɪ k] （色）
i	[s i] （絲）
ɛ	[s ɛ] （借）
ʊ	[s ʊ ŋ] （鬆）
u	[f u] （夫）
p^h	[p^h a] （扒）
t^h	[t^h a] （他）
ts	[ts i] （之）
ts^h	[ts^h i] （痴）
tʃ^h	[tʃ^h y] （處）
k^h	[k^h a] (卡)
k^w	[k^w a] （瓜）
k^{wh}	[k^{wh} a] （誇）
y	[ʃ y] （書）
œ	[h œ] （靴）
a	[s a] （沙）
ɐ	[s ɐ p] （濕）
ɵ	[s ɵ t] （恤）
ɔ	[s ɔ] （梳）
ei	[h ei] （稀）
ɛu	[t ɛu] （投）
ai	[w ai] （威）
ɵy	[s ɵy] （衰）
ɐi	[s ɐi] （西）
ui	[f ui] （灰）
iu	[s iu] （燒）
ɐu	[s ɐu] （收）
au	[s au] （筲）
ɔi	[s ɔi] （鰓）
ou	[s ou] （鬚）

English phonemes	
IPA symbol	Example
p	[p aɪ] (pie)
m	[m aɪ] (my)
f	[f l aɪ] (fly)
t	[t aɪ] (tie)
tʃ	[tʃ ɪ n] (Chin)
n	[n ɛ t] (net)
s	[s æ t] (sat)
ʃ	[ʃ aɪ] (shy)
l	[l aɪ] (lie)
j	[j u] (you)
k	[k aɪ t] (kite)
ŋ	[h æ ŋ] (hang)
w	[w aɪ] (why)
h	[h aɪ] (high)
ɪ	[b ɪ d] (bid)
i	[b i t] (beat)
ɛ	[b ɛ d] (bed)
ʊ	[g ʊ d] (good)
u	[b u t] (boot)
b	[b aɪ] (buy)
v	[v aɪ] (vie)
θ	[θ ɪ ŋ] (thing)
ð	[ð e ɪ] (they)
d	[d aɪ] (die)
z	[z u] (zoo)
ɹ	[ɹ ɛ n t] (rent)
dʒ	[p e ɪ dʒ] (page)
ʒ	[æ ʒ ɚ] (azure)
g	[g aɪ] (guy)
e	[b e ɪ t] (bait)
æ	[b æ d] (bad)
ɚ	[b ɚ d] (bird)
o	[b o t] (boat)
ɑ	[p ɑ d] (pod)
ʌ	[b ʌ d] (bud)
aʊ	[k aʊ] (cow)
aɪ	[b aɪ] (buy)
ɔɪ	[b ɔɪ] (boy)

In this section, we use IPA symbols to facilitate an intuitive comparison between Cantonese and English. Language-specific phonemic symbols have been commonly used in monolingual ASR research, for examples, Pinyin for Mandarin, Jyut-Ping for Cantonese (LSHK, 1997), and ARPABET for American English (Shoup, 1980). In Section 4, where phoneme-based acoustic modeling is discussed, we will use Jyut-Ping and ARPABET for monolingual Cantonese and English respectively, and a combination of them for code-mixing speech.

2.2 Properties of Cantonese-English Code-mixing Speech

Table 3 gives an example of Cantonese-English code-mixing sentence spoken in Hong Kong. It contains an English segment with one word. In this case, the English word is used as a substitute for its Chinese equivalent. The grammatical structure is totally that of Cantonese. In our application, the mother tongue of the speaker is Cantonese, *i.e.*, the matrix language. It is inevitable that the embedded English words carry Cantonese accent to certain extent. In many cases, the syllable structure of an English word changes to follow the structure of legitimate Cantonese syllables (Li, 1996). Such changes usually involve phone insertions or deletions. For example, the second consonant in a CCVC syllable of English may be softened, *e.g.*, the word "plan" in the example of Table 3 is pronounced as /p æ n/ instead of /p l æ n/ by many Cantonese speakers. A monosyllabic word with the CVCC structure may become disyllabic by inserting a vowel at the end, *e.g.*, /f æ n z/ ("fans") becoming /f æ n s ɪ/. It is also noted that the final stop consonant in an English word tends to be softened or dropped, *e.g.*, /t ɛ s t/ ("test") becoming /t ɛ s/. This is related to the fact that the stop coda of a Cantonese syllable is unreleased (Ching *et al.,* 2006). In addition to phone insertion and deletion, there also exist phone changes in Cantonese-accented English. That is, an English phoneme that is not found in Cantonese is replaced by a Cantonese phoneme that people consider to sound similar. For example, /θ r i/ ("three") becomes /f r i/ in Cantonese-accented English. Cantonese speakers in Hong Kong sometimes create a Cantonese pronunciation for an English word. For example, the word "file" (/f aɪ l/) is transliterated as /f aɪ l o/ (快佬 in written form). It is not a straightforward decision whether such a word should be treated as English or Cantonese. This is known as "lexical borrowing" (Chan, 1992).

In conclusion, English words in a code-mixing utterance must not be treated as being the same as those in a monolingual utterance from a native English speaker. For the design of ASR systems, special considerations are needed in acoustic modeling and lexicon construction.

Code-mixing occurs less frequently in read-style speech than in casual conversational speech. There exist many pronunciation variations in casual Cantonese speech, especially when the speaking rate is fast. Speakers may not follow strictly the pronunciations as specified in a standard dictionary. In the example of Table 3, the initial consonant /**n**/ of the first syllable is commonly pronounced as /**l**/ by the younger generation. Syllable fusion is often seen in fast speech, *i.e.*, the initial consonant of the second syllable of a disyllabic word tends to be omitted or changed (Kam, 2003; Wong, 2004).

Table 3. An example of a Cantonese-English code-mixing sentence

<table>
<tr><th colspan="5">Code-mixing speech</th></tr>
<tr><td>你哋
You (plural)</td><td>plan
plan</td><td>咗
already</td><td>行程
schedule</td><td>未 ?
or not</td></tr>
<tr><th colspan="5">Transcription according to standard pronunciation dictionary</th></tr>
<tr><td>/nei/ /tei/</td><td>/p l æ n/</td><td>/ts ɔ/</td><td>/h ɐŋ/ /ts^h ɪ ŋ/</td><td>/m ei/</td></tr>
<tr><th colspan="5">Transcription according to typical pronunciation in code-mixing speech</th></tr>
<tr><td>/lei/ /tei/</td><td>/p æ n/</td><td>/ts ɔ/</td><td>/h ɐŋ/ /ts^h ɪ ŋ/</td><td>/m ei/</td></tr>
<tr><th colspan="5">English translation</th></tr>
<tr><td colspan="5">Have you planned your schedule already?</td></tr>
</table>

2.3 Problems and Difficulties in Code-mixing Speech Recognition

Large-vocabulary continuous speech recognition (LVCSR) systems deal with fluently spoken speech with a vocabulary of thousands of words or more (Gauvain & Lamel, 2000). As shown in Figure 1, the key components of a state-of-the-art LVCSR system are acoustic models, pronunciation dictionary, and language models (Huang *et al.*, 2001). The acoustic models are a set of hidden Markov models (HMMs) that characterize the statistical variation of the input speech features. Each HMM represents a specific sub-word unit such as a phoneme. The pronunciation dictionary and language models are used to define and constrain the ways in which the sub-word units can be concatenated to form words and sentences.

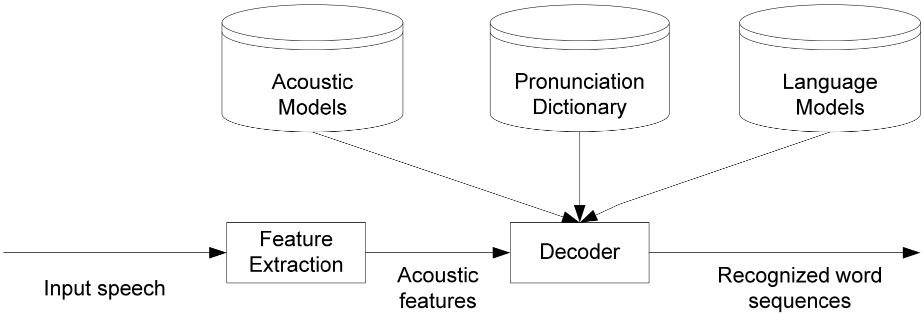

Figure 1. The flow diagram of an LVCSR system

For code-mixing speech recognition, the input utterance contains both Cantonese and English. Thus, the acoustic models are expected to cover all possible phonemes in the two languages. There are two possible approaches: (1) monolingual modeling with two separate sets of language-specific models; (2) cross-lingual modeling with some of the phoneme models shared between the two languages. Monolingual modeling has the advantage of preserving the language-specific characteristics and is most effective for monolingual speech from native speakers (Schultz & Waibel, 1998). In code-mixing speech where the English words are Cantonese-accented, an English phoneme tends to resemble or even become identical to a Cantonese counterpart. In this case, we may treat them as the same phoneme and establish a cross-lingual model to represent it. As shown in Table 2, Cantonese and English have a number of phonemes that are phonetically identical or similar to each other. The degree of similarity varies. In principle, cross-lingual modeling can be applied to those highly similar phonemes, while language-specific models would be more appropriate if the phonetic variation is relatively large. In Section 4, we are going to compare the effectiveness of cross-lingual and mono-lingual acoustic modeling and try to establish an optimal phoneme set for code-mixing speech recognition.

The pronunciation dictionary for code-mixing speech recognition is a mixture of English and Cantonese words. Each word may correspond to multiple pronunciations, which are represented in the form of phoneme sequences. Due to the effect of the Cantonese accent, the English words in code-mixing speech are subject to severe pronunciation variation as compared to those in standard English by native speakers. It is essential to reflect such variation in the pronunciation dictionary. On the other hand, as discussed in Section 2.2, the common pronunciation variations in spoken Cantonese should also be included.

In our application, the most common type of code-mixing is where one or more Cantonese words in the utterance being replaced by the English equivalent (Chan, 1992). The grammatical structure of code-mixing sentences is based largely on that of monolingual Cantonese. Word n-gram is by far the most commonly used technique for language modeling in LVCSR. To train a set of good n-gram models, a large number of spoken materials in computer-processable text format are needed. This presents a great challenge to our research since it is difficult in practice to find such materials for code-mixing speech. For the training of acoustic models, we need a large amount of code-mixing speech data. Development of speech and text corpora is therefore an important part of our work.

3. Development of Code-mixing Speech Corpus

In this section, the design, collection, and annotation of a Cantonese-English code-mixing speech corpus, named CUMIX, are described (Chan *et al.,* 2005). CUMIX is intended mainly for acoustic modeling for large-vocabulary speech recognition.

3.1 Corpus Design

There are three different types of utterances in CUMIX:

1. Cantonese-English code-mixing utterances (CM)
2. Monolingual Cantonese utterances (MC)
3. Monolingual English words and phrases (ME)

The CM utterances represent the typical code-mixing speech being dealt with in our application. There are practical difficulties in designing the content of code-mixing sentences. This is because spoken Cantonese is considered as a colloquial language that mainstream written publications do not use. Although the grammar of spoken Cantonese is similar to that of standard written Chinese, the lexical preference is quite different. An example pair of spoken Cantonese and written Cantonese sentences is shown in Table 4. Spoken Cantonese rarely appears in published text materials. Thus, text materials that involve code-mixing of spoken Cantonese and English are very limited.

Table 4. Comparison of spoken Cantonese and standard Chinese

Written Chinese:	你	吃過	午飯	了嗎?
Spoken Cantonese:	你	食咗	晏	未?
English translation (word by word):	You	eaten	lunch	or not?
English translation (whole sentence):	Have you had lunch?			

The design of CM sentences in CUMIX was based on a few local newspapers and online resources, including newsgroups and online diaries. We also consulted previous linguistic studies on Cantonese-English code-mixing. In Chan (1992), about 600 code-mixing sentences were analyzed. In 80% of the cases, the English segment contains a single word. The percentage distribution of nouns, verbs, and adjectives/adverbs are 43%, 24%, and 13%, respectively. There are very few cases involving prepositions and conjunctions. We try to follow these distributions in our corpus design.

A total of 3167 distinct code-mixing sentences were manually designed. Each sentence has exactly one English segment, which may contain one or more words. There are a total of 1097 distinct English segments. Each of them may appear more than once in the corpus, and if it does, the Cantonese contents of the respective sentences are different. The selected English words/phrases are commonly found in code-mixing speech and cover different part-of-speech categories.

The monolingual Cantonese sentences (MC) are identical to the CM sentences except that the English segments are replaced by the corresponding Cantonese words. The number of distinct MC sentences is smaller than that of CM ones because some of the English segments do not have Cantonese equivalents. Table 5 gives an example pair of CM and MC sentences.

In this example, the code-switched word “bonus” is replaced by the Cantonese word “花紅”.

Table 5. A CM sentence and the corresponding MC sentence

CM sentence:	我覺得今年有 bonus 嘅機會好渺茫。
MC sentence:	我覺得今年有 花紅 嘅機會好渺茫。
English translation:	I believe that it is very unlikely to have a bonus this year.

We also need English speech data for acoustic modeling of the English segments. Existing English databases like TIMIT and WSJ (Garofolo *et al.,* 1993; Lamel *et al.,* 1986; Paul & Baker, 1992) do not serve the purpose as they cannot reflect the phonetic and phonological properties of Cantonese-accented English. The amount of English speech data in the CM utterances is very limited. Thus, monolingual English utterances (ME) were also included as part of CUMIX to enrich the training data for the English acoustic models. The ME utterances contain English words and phrases, numbers and letters, which are most commonly used in Cantonese-English code-mixing speech.

3.2 Data Collection & Verification

The speech data in CUMIX were recorded from 34 male and 40 female native Cantonese speakers. Most of the speakers were university students. The average age was 22. The recording was carried out in a quiet room using a high-quality headset microphone. Each speaker was given a list of pre-selected sentences or phrases. He/she was requested to read each sentence fluently and naturally at a normal speaking rate. The speaker was also advised to adopt the pronunciations that they use in daily life.

Each recorded utterance was checked manually. The instants of language switching were marked. For those containing undesirable content or recording artifacts, the speakers were requested to record them again or the utterances were simply discarded. Each verified utterance is accompanied by an orthographic transcription, which is a sequence of Chinese characters with English words inserted in-between. In addition, the Cantonese pronunciations of the characters were also provided in the form of Jyut-Ping symbols.

3.3 Corpus Organization

Based on the usage, the utterances were organized into two parts, namely training data and test data. The training data set includes utterances from 20 male and 20 female speakers. Each speaker has 200 CM utterances and 100 ME utterances. Test data are intended for performance evaluation of the code-mixing speech recognition system and language boundary detection algorithms. There are 14 male and 20 female speakers in the test data. Each of them has 120 CM utterances and 90 MC utterances. Among the 34 test speakers, 5 males and 5 females were reserved as development data, which is intended for the tuning of various weighting parameters and thresholds in the system design. Table 6 gives a summary of the CUMIX corpus.

Table 6. A summary of CUMIX

		Training data	Test data
		20 male, 20 female	14 male, 20 female
CM	Duration:	7.5 hours	4.25 hours
	Duration of English segments:	1.13 hours	0.57 hours
	Total no. of utterances:	8000	3740
	No. of unique sentences:	2087	2256
	No. of unique English segments:	1047	1069
MC	Duration:		2.75 hours
	Total no. of utterances:		3060
	No. of unique sentences:		1742
ME	Duration:	1.5 hours	
	Total no. of utterances:	4000	
	No. of unique sentences:	1000	

4. Acoustic Modeling

This part of research aims at designing an appropriate phoneme inventory for acoustic modeling of Cantonese-English code-mixing speech. It is expected that some of the phoneme models are language-specific and the others are shared between Cantonese and English. Speech recognition experiments are carried out to evaluate the performances of three different sets of acoustic models in terms of syllable and word accuracy. In addition to CUMIX, two large-scale monolingual speech databases, namely TIMIT and CUSENT, are involved. CUSENT is a read-speech database developed for Cantonese LVCSR applications (Lee *et al.,* 2002). TIMIT is a phonetically balanced speech database of American English with hundreds of speakers (Garofolo *et al.,* 1993).

Table 7 explains the three sets of acoustic models, which are denoted by ML_A, ML_B, and CL, respectively. ML_A and ML_B are language-dependent phoneme models, in which Cantonese and English phonemes are separated despite the fact that some of them are phonetically similar. There are 56 Cantonese phonemes as listed in Table 8. They are adequate to compose all legitimate syllables of Cantonese. The English phoneme set has 39 elements as shown in Table 9. This phoneme set has been the most widely used in previous research (Lee & Hon, 1989). The difference between ML_A and ML_B is that they are trained with different training data as shown in Table 7.

CL is a set of cross-lingual models, designed to accommodate both Cantonese and English. As the matrix language, all Cantonese phonemes are included in the cross-lingual phoneme set. The English phonemes are divided into two parts. Phonemes that are unique to English are modeled separately, while the others are treated as Cantonese phonemes. In our work, the merging between English and Cantonese phonemes is based largely on phonetic knowledge (Chan, 2005). Due to the Cantonese accents, a number of English phonemes in the code-mixing speech are found to be sharable with Cantonese. It is also practically preferable to reduce the total number of phonemes as far as possible to facilitate effective utilization of training data. As a result, a total of 70 phonemes are selected for CL (Chan *et al.,* 2006). They are listed in Table 10. In addition to the 56 Cantonese phonemes in Table 8, a number of Cantonese diphthongs that have English equivalents are included. There are only 7 English-specific phonemes, while the others are mapped to some Cantonese equivalents.

Table 7. Different acoustic models being evaluated

Model	Phoneme inventory	Training data
ML_A	39 English phonemes 56 Cantonese phonemes	English: TIMIT Cantonese: CUSENT
ML_B	39 English phonemes 56 Cantonese phonemes	English: CUMIX Cantonese: CUSENT & CUMIX
CL	70 Cross-lingual phonemes	English: CUMIX Cantonese: CUSENT & CUMIX

Table 8. 56 Cantonese phonemes for monolingual modeling (ML_A & ML_B). Jyut-Ping symbols are used. “f-“ represents a syllable-initial consonant and “-m” represents a syllable coda. “k-/kw-” means that the two initial consonants are merged as one. “s-(yu)” represents a variant of “s-” when followed by the vowel “yu”.

Consonant	f-, h-, k-/kw-, g-/gw-, l-/n-, m, m-, -m, -n, ng, ng-, -ng, null, b-, p-, s-, s-(yu), z-, z-(yu), c-, c-(yu), d-, t-, w-, j-
Vowel	a, aa, o, e, eo, i, i(ng), oe, u, u(ng), yu
Vowel-stop	ap, at, ak, aap, aat, aak, ep, et, ek, ut, uk, yut, ip, it, ik, op, ot, ok, eot, oek

Table 9. English phonemes for monolingual modeling (ML_A & ML_B). APRABET symbols are used to label the phonemes.

Consonant	dh, th, f, v, w, z, zh, s, sh, t, d, b, p, ch, g, h, jh, k, l, m, n, ng, y, r
Vowel	aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw

Table 10. Phonemes for cross-lingual modeling (CL). English-specific phonemes start with the prefix "E_" and are labeled with ARPABET symbols.

Consonant (30)	f-, h-, k-/kw-, g-/gw-, l-/n-, m, m-, -m, -n, ng, ng-, -ng, null, b-, p-, s-, s-(yu), z-, z-(yu), c-, c-(yu), d-, t-, w-, j-, E_t, E_d, E_k, E_r, E_z
Vowel/diphthong (20)	a, aa, o, e, eo, i, i(ng), oe, u, u(ng), yu, iu, aai, ai, au, ou, oi, ei, E_ah, E_el
Vowel-stop (20)	ap, at, ak, aap, aat, aak, ep, et, ek, ut, uk, yut, ip, it, ik, op, ot, ok, eot, oek

The English phoneme models in ML_A are trained with TIMIT, and the Cantonese models are trained with CUSENT. The English words in TIMIT sentences are transcribed into phoneme sequences based on the CMU pronunciation dictionary (CMU). The Cantonese syllables in CUSENT utterances are transcribed into phoneme sequences using a standard Cantonese pronunciation dictionary (LSHK, 1997). All training data are assumed to follow the standard pronunciations.

For ML_B, the English phoneme models are trained with the code-switched English segments in the CM and ME utterances of CUMIX. The Cantonese phoneme models are trained with CUSENT and the Cantonese part of CUMIX. Moreover, the pronunciation dictionaries used for transcribing the utterances include not only standard English but also Cantonese-accented English and common pronunciation variants of Cantonese syllables. Thus, there may exist multiple pronunciations for a lexical entry. For each of the possible pronunciations, the acoustic likelihood of the word or syllable segment is computed. The pronunciation with the highest likelihood is adopted for the training of ML_B.

For CL, we use the same training data as for ML_B. We also use the same transcriptions as determined for ML_B except that the language-dependent phoneme symbols are converted into the cross-lingual phoneme symbols in Table 10.

The effectiveness of ML_A, ML_B, and CL are evaluated by syllable/word recognition experiments. The test data include the CM and the MC test utterances of CUMIX. The acoustic feature vector has 39 components: 13 MFCC and their first and second-order time derivatives. All phoneme models are context-dependent triphone HMMs. Each model consists of three emitting states, each of which is represented by a mixture of Gaussian density functions. States in models are clustered and tied using a decision-tree based technique with pre-set phonetic questions. ML_A and ML_B use 16 Gaussian components per state, while CL has 32 Gaussian components. The grammar network used for recognizing CM utterances is illustrated in Figure 2. For MC utterances, the recognition network is simplified into a syllable loop.

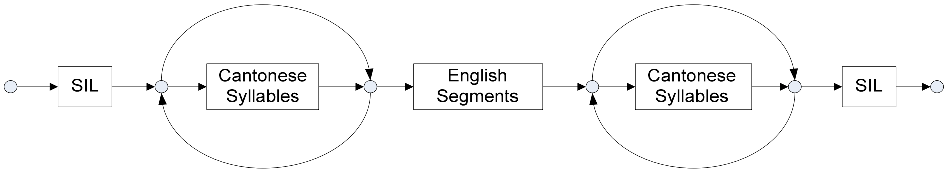

Figure 2. Grammar network for syllable/word recognition of code-mixing speech

The recognition performance is measured in terms of syllable accuracy for Cantonese and word accuracy for English. The test results are given in Figure 3. For code-switched English words, ML_A attains a very low accuracy of 18.9%. This confirms that Cantonese-accented English is very different from the native American English found in TIMIT. ML_B improves greatly in recognizing English words due to better matched training. Nevertheless, the accuracy of 40.5% is still on the low side because of the limited amount of training data and the language-dependent nature of the models. The English words in CUMIX carry Cantonese accents, such that some of the English phoneme models are very close to certain Cantonese phoneme models. In other words, similar acoustic features are captured by two different models. Hence, the confusion between English words and Cantonese syllables tends to increase. The Cantonese syllables are easily misrecognized as English words, and vice-versa. This also explains why the performance of ML_B in recognizing Cantonese syllables declines.

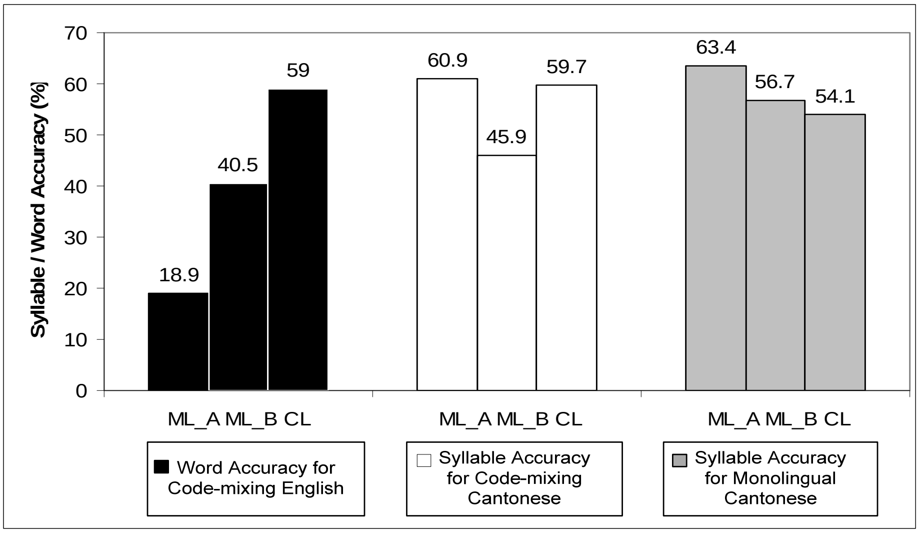

Figure 3. Syllable/word accuracy of the three acoustic models

For Cantonese speech in the code-mixing utterances (CM), the recognition accuracies attained by ML_A and ML_B are 60.9% and 45.9% respectively. The poor performance of ML_B is related to the use of language-dependent models as discussed above. The performance difference between ML_A and ML_B for monolingual Cantonese utterances (MC) is not as significant as for the CM utterances. This is because the grammar network used for MC utterances does not include an English segment, and therefore there should be no recognition error caused by the confusion between similar Cantonese and English phonemes.

CL uses a large number of shared phoneme models between English and Cantonese. It attains the best recognition accuracy of 59% for the embedded English words, and at the same time, it maintains a reasonable performance on Cantonese. It is believed that the existing design of cross-lingual models can be improved further with more understanding about the phonetic variation in code-mixing speech. More training data will also be helpful.

5. Language Modeling

5.1 Collection and Selection of Text Data

There are practical difficulties in collecting a large amount of text material to facilitate statistical language modeling for Cantonese-English code-mixing speech. Cantonese is a spoken dialect; many colloquial Cantonese words do not have a standard written form. In addition, written Cantonese is neither taught in schools nor recommended for official and documentary usage. Nevertheless, a limited amount of Cantonese text data can be found in certain columns of local newspapers, magazines, advertisements, and online articles (Snow, 2004). On the other hand, code-mixing is a domain-specific phenomenon. It is found in the discourses that involve contemporary and cross-cultural issues, *e.g.*, computer, business, fashion, food, and showbiz (Li, 1996). In our study, Cantonese text data are selected from three major sources, namely newspaper, magazines, and online diaries. Preliminary manual inspection was done to identify the sections or columns that are highly likely to contain code-mixing text. A total of 28 Chinese characters that are frequently used in spoken Cantonese but rarely used in standard Chinese were identified, *e.g.*, 嘅, 嘢, 咁 (Snow, 2004). Articles that contain these characters were considered to be written in Cantonese. As a result, a text database with 6.8 million characters was compiled. There are about 4600 distinct Chinese characters and 4200 distinct English segments in the database. About 10% of these English segments are included in the CUMIX utterances.

5.2 Training of Language Models

The text data were used to train character tri-grams. Four different models were trained:

CAN_LM: mono-lingual Cantonese language model;

CM_LM: code-mixing language model;

CLASS_LM: class-based language model;

TRANS_LM: translation-based language model.

For CAN_LM, all English words were removed from the training text. They were considered as out-of-vocabulary (OOV) words during the evaluation. OOV words are assigned zero probability so that they may be missed in recognition. For CM_LM, all code-switched segments in the training text were mapped to the same word ID during the training process, no matter whether the words were found in the training text or not. By doing so, the likelihood of English segments is made much higher than that of the Cantonese characters, thus, Cantonese words may be easily misrecognized as English words. In CLASS_LM, code-switched segments were divided into 15 classes according to their parts of speech (POS) or meanings. Most of the classes were for nouns. TRANS_LM involves English-to-Cantonese translation, by which code-switched segments are translated into their Cantonese equivalents. Nevertheless, since not all of the code-switched terms have Cantonese equivalents, the POS classes being used in CLASS_LM were considered as well.

The language models were evaluated in the phonetic-to-text (PTT) conversion task. Assuming that the true phonetic transcription is known, language models were used to determine the word sequence that best matched the transcription. For Chinese languages, PTT conversion is often formulated as a problem of syllable-to-character or Pinyin-to-text conversion. Statistical language models have proven to be very effective (Gao *et al.,* 2002). In our study, PTT conversion was treated as a sub-task of decoding for speech recognition. The proposed code-mixing speech recognition system employs a two-pass decoding algorithm (see Section 7 for details). The first pass generates a syllable/word lattice using acoustic models and bilingual dictionary. Language models are used in the second pass to decode the Chinese character sequence. PTT conversion can be done by skipping the first pass and using the true syllable-level transcription to replace the hypothesized syllable lattice. In this way, the effectiveness of language models can be assessed. The true syllable transcription of the CM test utterances is used as the input. The PTT conversion accuracy attained by different language models is given in Table 11.

Table 11. Phonetic-to-text conversion rate by different language models

Language model	PTT conversion rate (character accuracy)
Monolingual Cantonese (CAN_LM)	88.8%
Code-mixing (CS_LM)	89.3%
Class-based (CLASS_LM)	91.5%
Translation-based (TRANS_LM)	86.1%

The four language models are close to each other in performance because their differences are mainly on the code-switched segments. The translation approach (TRANS_LM) achieves the lowest PTT conversion rate. This is due to some of the translated Cantonese characters not appearing in the character list of the original Cantonese language models. This leads to the n-gram probabilities that are related to these characters being very low in TRANS_LM. The low likelihood affects the decision on the neighboring characters and leads to degradation of the overall conversion rate. Moreover, the code-switched segments are translated into Cantonese, and each translated term may contain more than one character. This causes a discrepancy in the computed values of the PTT conversion rate.

6. Language Boundary Detection

Language identification (LID) is an important process in a multilingual speech recognition system (Ma *et al.,* 2007). The language identity information allows the use of two monolingual recognizers. However, the LID for recognizing code-mixing speech is not straightforward mainly because the speech segments that can be used for decisions are relatively short. For code-mixing speech, LID can be considered as a problem of language boundary detection (LBD). We consider two approaches below (Chan *et al.,* 2006).

6.1 LBD based on syllable bigram

The syllable bigram probability of Cantonese is defined as the probability that a specific syllable pair occurs. In our study, these probabilities were computed from a transcribed Cantonese text database. In a code-mixing utterance, the Cantonese part is expected to have high syllable bigram probability, while the embedded English segments have relatively low syllable bigram probability, because of the mismatch in phonological and lexical properties. We use a Cantonese syllable recognizer based on the cross-lingual acoustic model CL as described in Section 4. For each pair of adjacent syllables in the recognized syllable sequence, the syllable bigram is retrieved. If the probability is higher than a threshold, this syllable-pair segment is considered to be Cantonese; otherwise, it is English or at the code-mixing boundary. It is possible that more than one English segment is detected within an utterance. Under the assumption that each utterance consists of exactly one English segment, we need to select one of the hypotheses. Our current strategy is to select the segment with the longest duration. On the other hand, if no English segment is found, the threshold is increased until the English segment includes at least one syllable.

To evaluate the performance of an LBD algorithm, the detected boundaries of a language segment are compared to the true boundaries. If the detection errors on both sides of the segment exceed a threshold, an LBD error is recorded. In this study, the threshold was set to 0.3 second. With the syllable bigram based detection algorithm, an LBD accuracy of 65.9%

was attained for the CM test utterances.

6.2 LBD based on Syllable Lattice

This approach makes use of the syllable/word lattice generated by a bilingual speech recognizer, which will be described in the next section. Syllable lattice is a compact representation of recognition output, which covers not only the best syllable sequence but also other possible alternatives. The lattice produced by our system contains Cantonese syllable units and English word/phrase units. English words/phrases generally have longer duration than Cantonese syllables since they may contain multiple syllables. The English segment with the longest duration in the lattice is most likely to indicate a correct recognition result, and the start and end time of the segment are taken as the language boundaries. With a properly selected insertion penalty, the LBD accuracy for CM test utterances was 82.3%.

7. A Code-mixing Speech Recognition System

7.1 System Overview and Decoding Algorithm

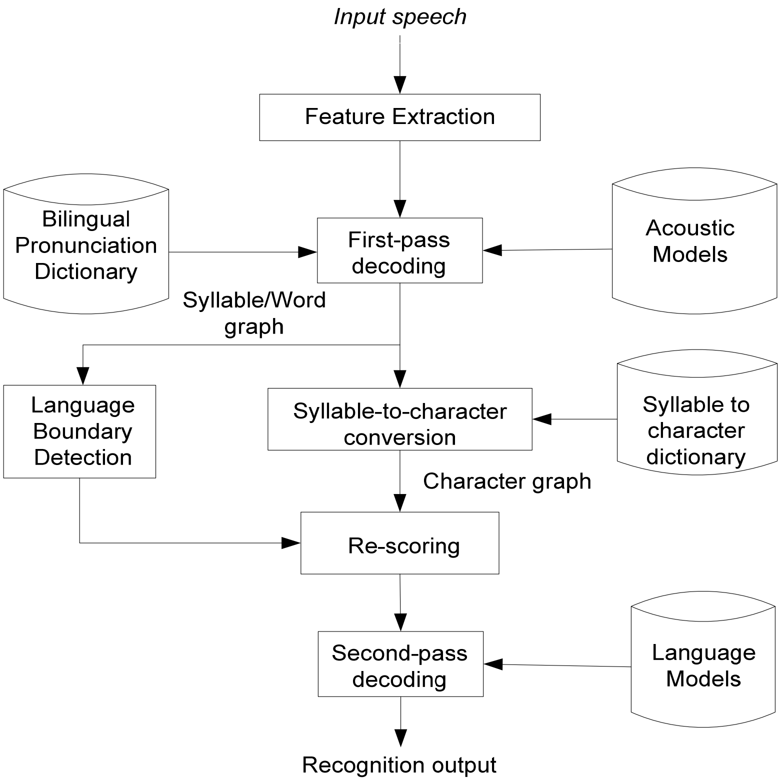

Figure 4. The proposed code-mixing speech recognition system

A code-mixing speech recognition system was developed as shown in Figure 4. It consists of the cross-lingual acoustic models, the bilingual pronunciation dictionary, and the class-based language models as described in previous sections. It is assumed that the input utterance is either code-mixing speech with exactly one English segment, or monolingual Cantonese speech. The decoding algorithm is implemented with the HTK Toolkits (Young *et al.,* 2001). It consists of two passes as described below.

First pass

In the first pass, the cross-lingual acoustic models and the bilingual pronunciation dictionary are used to construct a recognition network as shown in Figure 2. In the case where the input utterance is monolingual Cantonese, the recognition network is simplified into a syllable loop. Language models are not involved at this stage. The recognition network represents all possible hypotheses, from which the most likely ones are to be determined. The first-pass decoding is based on a token-passing algorithm. Each token refers to a partial hypothesis starting from the first frame of the utterance. At each time step, a feature vector is taken up and the existing tokens are extended through the HMM states in the recognition network. If there are many competing tokens at a network node, only the best N tokens are kept and the others are discarded. In this way, a syllable/word graph is generated as a compact representation of multiple hypotheses. The basic elements of the graph are nodes and arcs. Each arc represents a hypothesized Cantonese syllable or a hypothesized English word/phrase. It records the acoustic likelihood, the start time, and end time of the syllable or words/phrases. An example of mixed syllable/word graphs is shown in Figure 5.

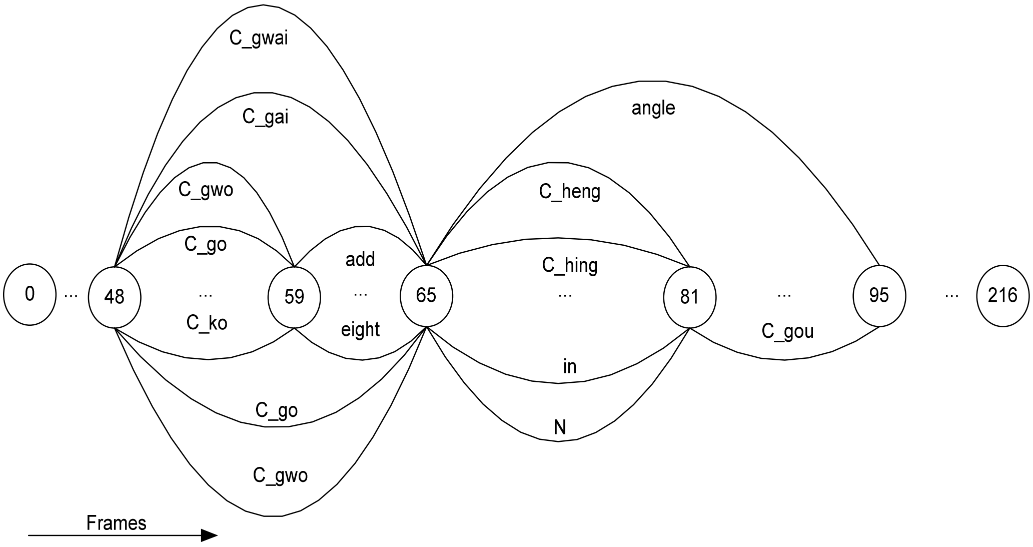

Figure 5. An example of mixed syllable/word graphs

<u>Second pass</u>

In the second pass, the most likely code-mixing sentence is determined from the syllable/word graph. In addition to the acoustic likelihoods, language boundary information and language models are utilized in the search process. Firstly, the language boundary information is integrated to the syllable/word lattice by modifying the acoustic likelihood of hypothesized words. If a hypothesized word is in the same language as the recognized language, the acoustic likelihood is increased by a pre-determined value; otherwise, it is decreased by the same value. The optimal value of this bonus/penalty score is derived from development data. Secondly, the modified acoustic scores are integrated with the language model scores to form a character lattice. The hypothesized syllables in the graph are mapped to Chinese characters using a pronunciation dictionary (LSHK, 1997). Since a Cantonese syllable may correspond to more than one Chinese character, the resulting character graph is in fact an expanded version of the syllable graph. The English words/phrases in the graph remain untouched. In the word graph, the posterior probability of a hypothesized word can be computed by summing the posterior probabilities of all sentence hypotheses that share the word segment w at the same time interval. In Soong *et al.* (2004), the generalized word posterior probability (GWPP) was formulated mainly to deal with the inconsistent dynamic ranges of acoustic models and language models, and with the alignment ambiguities between different sentence hypotheses. The effectiveness of GWPP has been demonstrated in Cantonese large-vocabulary continuous speech recognition (Qian *et al.,* 2006).

Let w denote a hypothesized word or syllable in the graph, with the start time s and end time t. The GWPP of w during the time interval $[s,t]$ is calculated from all word strings that contain w with a time interval overlapping with $[s,t]$, *i.e.*,

$$P([w;s,t]\,\big|x_1^T) = \sum_{\substack{W^M,\forall M\\ \exists n,1\le n\le M,\text{ s.t.}\\ w_n=w,\text{ and}\\ [s_n,t_n]\cap[s,t]\neq\Phi}} \frac{\prod\limits_{\text{m=1}}^{M} P^{\alpha}(x_{s_m}^{t_m}\,\big|w_m)\cdot P^{\beta}(w_m\,\big|w_1^{m-1})}{P(x_1^T)} \quad , \qquad (1)$$

where $W^M=\{[w_1;s_1,t_1],[w_2;s_2,t_2],\ldots,[w_M;s_M,t_M]\}$ denotes a specific word string that contains M words, and $[w_{n;}s_n,t_n]$ refers to the *n*th word in the string, which starts at time s_n and ends at t_n. The conditions of $w_n=w$ and $[s,t]\cup[s_n,t_n]\neq\Phi$ mean that the hypothesized word appears in this word string over approximately the same time interval. $P(x_{s_m}^{t_m}\,|w_m)$ and $P(w_m\,\big|w_1^{m-1})$ denote respectively the acoustic model scores and the language model scores. The prior probability $P(x_1^T)$ can be calculated by summing up all forward strings probabilities or backward string probabilities in the word graph. The weighting factors α and β are jointly optimized by using a held-out set of development data with a goal to achieve the minimum word error rate.

7.2 Experimental Results

The performance of the code-mixing speech recognition system in Figure 4 was evaluated using the CM and MC test utterances. For the CM utterances, the character accuracy was measured for the Cantonese part and the word accuracy is measured for the embedded English segments. From the development data in CUMIX (see Section 3.3), the best values α and β were found to be 0.009 and 1.1 respectively. This leads to an overall accuracy of 55.1% for the development utterances.

Without the use of language boundary detection, the overall recognition accuracy for CM and MC utterances were 55.3% and 50.3%, respectively, when the class-based language models CLASS_LM were used. The detailed results are given in Table 12.

Table 12. Recognition accuracy without using language boundary information

	Overall accuracy	Cantonese Character accuracy	English Word accuracy
CM test utterances	55.3%	56.0%	48.4%
MC utterances	50.3%	50.3%	

We also attempted to incorporate the detected language boundaries into the recognition process. Table 13 compares the effectiveness of the two LBD approaches described in Section 6. With LBD based on syllable bigram, the overall recognition accuracy increases from 55.3% to 57.0%. For the syllable-lattice based LBD, although the overall accuracy does not increase significantly, there is a noticeable improvement on the recognition accuracy for the English words. Among the recognition errors on English words, 39.0% of them are deletion errors, while 44.2% are substitution errors. Deletion error means that no English word is found in the top-best hypothesis string. Substitution errors are mainly caused by incorrect language boundary thus the hypothesis English word and the reference English word have no or just very little overlap in time duration. For example, the word "evening" is mistakenly recognized as "even", and "around" became "round".

It was also noted that the English word accuracy could be improved to 81.1% if the true language boundaries are used in the recognition process. It is believed that the recognition performance can be improved, when better language boundary detection algorithms become available.

Table 13. Recognition accuracy attained with the incorporation of language boundary information. Only CM test utterances are used.

	Overall accuracy	Cantonese Character accuracy	English Word accuracy
Without LBD	55.3%	56.0%	48.4%
LBD based on syllable bigram	57.0%	57.6%	49.0%
LBD based on syllable lattice	56.0%	56.4%	53.0%

For Cantonese, the character accuracy was close to our expectation. The character accuracy (56.4%) was roughly equal to the syllable accuracy (59.7%) multiplied by the PTT conversion rate (91.5%).

8. Conclusion

Code-mixing speech recognition is a challenging problem. The difficulties are two-fold. Firstly, we have little understanding about this highly dynamic language phenomenon. Our study clearly reveals that code-mixing is not a simple insertion of one language into another. It comes with a lot of phonological, lexical, and grammatical variation with respect to monolingual speech spoken by native speakers. Unlike in monolingual speech recognition research, there are very few linguistic studies that can be consulted. We have to understand the problems by actually working on them. Secondly, it is practically difficult to collect sufficient code-mixing data for effective acoustic modeling and language modeling. The existing CUMIX database needs to be enhanced, especially in the amount of English speech.

We have shown that cross-lingual acoustic models are more appropriate than language-dependent models. The proposed cross-lingual models attain an overall recognition accuracy of nearly 60% for code-mixing utterances. To design a cross-lingual phoneme set, we need to measure the similarity between the phonemes of the two languages. Our current approach is based on phonetic knowledge. It can be improved further with comprehensive acoustic analysis of real speech data. For language modeling, grouping English words into classes seems to be inevitable due to data sparseness. The class-based language models were shown to be effective in code-mixing speech recognition.

Two different methods of language boundary detection have been evaluated. LBD based on syllable bigram exploits the phonological and lexical differences between Cantonese and English. LBD based on syllable lattice makes use of the intermediate result of speech recognition, which is more informative than the prior linguistic knowledge. Therefore, this method attains a significantly higher accuracy than the former one in language boundary detection.

A complete speech recognition system for Cantonese-English code-mixing speech has been developed. The two-pass search algorithm enables flexible integration of additional knowledge sources. The overall recognition accuracy for Cantonese syllables and English words in code-mixing utterances is 56.0%.

Acknowledgements

This work is partially supported by a RGC grant (No. 411306) and also a grant awarded by the Shun Hing Institute of Advanced Engineering.

References

Auer, P. (1998). *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge, London.

Chan, H. S. (1992). *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*. M.A. Thesis, The Chinese University of Hong Kong, Hong Kong.

Chan, J. Y. C., Ching, P. C., Lee, T. and Meng, H. (2004). Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong, 293-296.

Chan, J. Y. C. (2005). *Automatic Speech Recognition of Cantonese-English Code-Mixing Utterances*, M.Phil. Thesis, the Chinese University of Hong Kong, Hong Kong.

Chan, J. Y. C., Ching, P. C. and Lee, T. (2005). Development of a Cantonese-English code-mixing speech corpus. In *Proceeding of Eurospeech*. Lisbon, Portugal, 1533-1536.

Chan, J. Y. C., Ching, P. C., Lee, T. and Cao, H. (2006). Automatic speech recognition of Cantonese-English code-mixing utterances. In *Proceeding of Interspeech (ICSLP)*. PA, USA, 113-116.

Chan, C.-M. (2004). Two types of code-switching in Taiwan. In *Proceeding of the 15th Sociolinguistics Symposium*. Newcastle, UK.

Ching, P. C., Lee, T., Lo, W. K. and Meng, H. (2006). Cantonese speech recognition and synthesis, In *Advances in Chinese Spoken Language Processing*, Lee, C.-H., Li, H., Lee, L.-S., Wang, R.-H., and Huo, Q., Eds. World Scientific Publishing, Singapore, 365-386.

Ching, P. C., Lee, T. and Zee, E. (1994). From phonology and acoustic properties to automatic recognition of Cantonese. In *Proceeding of International Symposium on Speech, Image Processing and Neural Networks*. Vol. 1. Hong Kong, 127-132.

Carnegie Mellon University (CMU). *The CMU Pronouncing Dictionary v0.6*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Gao, J., Goodman, J., Li, M. and Lee, K.-F. (2002), Toward a unified approach to statistical language modeling for Chinese. In *ACM Trans. on Asian Language Information Processing*, 1(1), 3-33.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1993). *DARRA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST Speech Disc1-1.1, NISTIR 4930.

Gauvain, J. L. and Lamel, L. (2000). Large-vocabulary continuous speech recognition: advances and applications. In *Proceeding of the IEEE*, 88(8), 1181-1200.

Gumperz, J. (1982). *Discourse Strategies*. Cambridge University Press, Cambridge, 59.

Halmari, H. (1997). *Government and Code-switching: Explaining American Finnish*. J. Benjamins, Amsterdam.

Huang, X., Acero, A. and Hon, H.-W. (2001). *Spoken Language Processing*. Prentice-Hall, New Jersey.

Kam, P. (2003). *Pronunciation Modeling for Cantonese Speech Recognition*. M.Phil. Thesis, The Chinese University of Hong Kong, Hong Kong.

Lamel, L. F., Kassel, R. H. and Seneff, S. (1986). Speech database development: design and analysis of the acoustic-phonetic corpus. In *Proceeding of DARPA Speech Recognition Workshop*. Palo Alto, 100-109.

Lee, K.-F. and Hon, H.-W. (1989). Speaker-independent phone recognition using hidden Markov models. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(11), 1641-1648.

Lee, T., Lo, W. K., Ching, P. C. and Meng, H. (2002). Spoken language resources for Cantonese speech processing. In *Speech Communication*, 36(3-4), 327-342.

Li, D. C. S. (1996). *Issues in Bilingualism and Biculturalism: a Hong Kong Case Study*. Peter Lang Publishing, New York.

Li, D. C. S. (2000). Cantonese-English code-switching research in Hong Kong: a Y2K review. In *World Englishes*, 19(3), 305-322.

Li, P. (1996). Spoken word recognition of code-switched words by Cantonese-English bilinguals. In *Journals of Memory and Language*, 35, 757-774.

Linguistic Society of Hong Kong (1997). *Jyut Ping Characters Table*. Linguistic Society of Hong Kong Press, Hong Kong.

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., and Hsu, C.-N. (2006). Speech recognition on code-switching among the Chinese dialects. In *Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. Toulouse, France, 1105-1108.

Ma, B., Li, H. and Tong, R. (2007). Spoken language recognition using ensemble classifiers. In *IEEE Trans. on Audio, Speech and Language Processing*, 15(7), 2053-2062.

Paul, D. B., and Baker, J. M. (1992). The design for the wall street journal-based CSR Corpus. In *Proceeding of Workshop on Speech and Natural Language*. New York, USA, 357-362.

Qian, Y. Soong, F. K. and Lee, T. (2005). Tone-enhanced generalized character posterior probability (GCPP) for Cantonese LVCSR," In *Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. Toulouse, France, 133-136.

Schultz, T. and Waibel, A. (1998). Language independent and language adaptive large vocabulary speech recognition. In *Proceeding of Interspeech (ICSLP)*. Sydney, Australia, 577-580.

Schultz, T. AND Kirchhoff, K. Eds. (2006). *Multilingual Speech Processing*. Elsevier Inc..

Shoup, J. E. (1980). Phonological Aspects of Speech Recognition. In *Trends in Speech Recognition,* Lea, W. A. Ed. Prentice-Hall, New York, 125-138.

Snow, D. (2004). *Cantonese as Written Language: the Growth of a Written Chinese Vernacular*. Hong Kong University Press, Hong Kong.

Soong. F. K., Lo, W. K., Nakamura, S. (2004). Optimal acoustic and language model weights for minimizing word verification errors. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong.

The International Phonetic Association (1999). *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Wester, M. (2003). Syllable classification using articulatory-acoustic features. In *Proceeding of Eurospeech*. Geneva, Switzerland, 233-236.

Wong, W. Y. (2004). Syllable fusion and speech rate in Hong Kong Cantonese. In *Proceeding of Speech Prosody*. Nara, Japan, 255-258.

Wu, C.-H., Chiu, Y.-H., Shia, C.-J. and Lin, C.-Y. (2006). Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. In *IEEE Transactions on Speech and Audio Processing*, 14, 266-276.

You, S.-R., Chien, S.-C., Hsu, C.-H., Chen, K.-S., Tu, J.-J., Lin, J.-S. and Chang, S.-C. (2004). Chinese-English mixed-lingual keyword spotting. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong, 237-240.

Young, S. et al. (2001). *The HTK Book (for HTK Version 3.1)*. Cambridge University, Cambridge.