Automatic speech recognition of Cantonese-English code-mixing utterances

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

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

Abstract

This paper describes our recent work on the development of a large-vocabulary, speaker-independent, continuous speech recognition system for Cantonese-English code-mixing utterances. The details of both acoustic modeling and language modeling will be discussed. For acoustic modeling, Cantonese accents in English words are handled by applying cross-lingual acoustic units, as well as modifications in pronunciation dictionary. Statistic language models are built from a small amount of text data, as there are many limitations on data collection. Language boundary detection based on language identification algorithms is applied, and it offers a slight improvement to the overall accuracy. The recognition accuracy for Chinese characters and English lexicons in the code-mixing utterances is 56.37% and 52.99%, respectively.

Index Terms: speech recognition, code-mixing, language detection

1. Introduction

Code-switching is defined as the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems [1]. Code-switching is common in many bilingual societies, and different combinations of languages are involved, for example, Spanish / English, French / Russian, Mandarin / Taiwanese, and so on [2]. To describe the phenomenon in greater detail according to the frequency of language switching, the terms (inter-sentential) code-switching and (intra-sentential) code-mixing are utilized. In Hong Kong, code-switching mainly occurs in the word level, hence the term code-mixing is usually preferred [3].

Hong Kong is a truly international city, and majority of the population is composed of Cantonese-English bilinguals. Code-mixing between Cantonese and English in speech as well as in written text is a common practice of local residents. The primary language of most residents in Hong Kong is Cantonese. In order to better describe meanings, feelings, and phenomena, people may switch to English when they are speaking Cantonese. The English involved is mainly comprised of single words or short phrases, and Cantonese accents are usually included. It means that Cantonese is the matrix language, while English is the embedded language [4]. In code-mixing utterances, the duration of the embedded language is relatively short. Hence, it is difficult to perform language identification. Together, the switching between languages and accents in the code-switch words will lead to difficulties in automatic speech recognition. The following is an example of code-mixing between Cantonese and English: 你 print 咗份功課未？(Have you printed the assignment?)

In order to study the effect of accents, the performance of monolingual and cross-lingual acoustic models is compared, and both monolingual and code-mixing speech corpora are involved. To handle accents in the code-switch words, the phonetic sequence of the English lexicons in the pronunciation dictionary is modified. Four different statistic language models are proposed in order to solve the problem on the lack of code-mixing training text data. Language boundary detection based on algorithms in language identification is studied and applied to the speech recognizer. The language boundary information is integrated to the speech recognition result by re-scoring the acoustic scores. Finally, the language model scores and the re-computed acoustic scores are integrated and re-weighted to obtain the Generalized Word Posterior Probability (GWPP) [5] for decoding.

2. Phonological structure of Cantonese and English

Cantonese and English come from two different language families, hence their contrasting phonological structures. 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 of C1VC2, where C1 and C2 are optional consonants, and V is either a simple vowel or a diphthong. All the Cantonese syllables are of the canonical forms V, CV, CVC, or VC. But English is an Indo-European language, the phonological structure of which is more complicated than Cantonese. In English discourse, over 80% of the syllables are of the canonical form of Cantonese, while the remaining are C, CC, CCV, VCC, CCCV, CCCVCC, and so on [6].

2.1. Cantonese accents in English words

In code-mixing utterances, the words in the embedded language may be pronounced with accents of the matrix language. Hence, for Cantonese-English code-mixing utterances, the syllable structures of English words are usually modified to those in Cantonese and become (C)V(C). Changes in syllable structures lead to phone insertion or phone deletion. For example, the second consonant of the CCVC structure is usually softened, and monosyllabic words with the CVCC structure will become CVC CV, if the third consonant is fricative. The final stop consonant will also be softened or dropped, since those in Cantonese are all unreleased [7].

Phone change is another effect of accents, which means that phones unique to English are usually pronounced as similar phones in Cantonese. Therefore, to recognize the accented English words, the clustering of acoustic units is proposed, and modifications in the phonetic sequence of English lexicons in the pronunciation dictionary are necessary.

2.2. Characteristics of spoken Cantonese

The official written language in Hong Kong is standard Chinese instead of Cantonese, since Cantonese is just a dialect. People use standard Chinese in formal reports and documentations, while spoken Cantonese in written form is used for soft news, advertisements, and other informal text. The lexicons and grammar of standard Chinese and spoken Cantonese are quite different, and there is no formal education on the written form of spoken Cantonese.

Some of the spoken Cantonese lexicons do not have a standard written form, as people may create the characters themselves or borrow characters with similar pronunciation. For example, the lexicon which means "yesterday" can be written in four different forms – 尋日, 噚日, 琴日, 撳日. Some of the colloquial words cannot be written in Chinese because no character has that pronunciation [8]. People may borrow English words with a similar sound for these words, for example, "烏 where" [wu wɛ] and "B Lee 巴啦" [pi li pa la]. These words should be distinguished from code-mixing, since we just borrow their sound instead of their meaning.

Apart from the written form, there are also many variations in pronunciation, since Romanization systems are not taught in school. Some of the consonants are easily confusing, and they are usually mispronounced as the other consonants. Table 1 shows some of the examples.

Syllable fusion is another type of pronunciation variation that mainly occurs in fast speech and common words [9]. The initial consonant is affected by the final consonant of the previous syllable. For example, 今日 [kɐm jɐt] may be pronounced as [kɐm mɐt], and 即刻 [tʃɪk hɐk] becomes [tʃɪk kɐk].

Table 1. Common confusing phones in Cantonese

The original phone	Realize as another phone
gʷɔk (國, 幗, 郭…)	gɔk(各, 角, 閣…)
nei (你, 尼, 餌…)	lei (李, 理, 里…)
hɐŋ (肯, 亨, 衡…)	hɐn (很, 痕, 恨…)
hɔk (殼, 學, 鶴…)	hɔt (喝, 渴, 褐…)

3. Acoustic modeling

Three speech corpora are involved in this research; they are the monolingual English corpus TIMIT, the monolingual Cantonese corpus CUSENT [10], and the Cantonese-English code-mixing corpus CUMIX [11]. TIMIT contains five hours of read speech from 630 speakers representing eight major dialect divisions of American English. CUSENT is a large collection of read Cantonese sentences that is designed to be phonetically rich. Sixty-eight native Cantonese speakers are involved, and the corpus size is 20 hours. CUMIX involves read speech in Cantonese-English code-mixing, monolingual Cantonese, and English lexicons with Cantonese accents. It contains nine hours of speech data from 40 speakers. Three sets of acoustic models are trained from these corpora as shown in Table 2.

All the acoustic models are tri-phones that depend on both the left and the right context. The language-dependent models are monolingual, which includes 39 English phones and 56 Cantonese phones. The English phones are ARPABET, and the pronunciation dictionary is based on the CMU pronunciation dictionary [12]. The Cantonese phones are mainly IPA phones, and the pronunciation dictionary is based on a Chinese syllabary pronunciation according to the Canton dialect (粵音韻彙). Since the stop consonants in Cantonese are unreleased, their effect are mainly reflected in the previous vowel. Thus, there will be acoustic units with a vowel-consonant structure.

English phones in model set A do not include Cantonese accents, since TIMIT is recorded by native speakers. In code-mixing, the code-switch words usually contain accents; thus, the accents should be included in the training data as well. Therefore, TIMIT is replaced by CUMIX in model sets B and C, and the difference between them is the language dependency of the models. In model set C, language-independent (cross-lingual) models are utilized. Similar phones of the two languages are clustered, and therefore, the total number of phones is reduced to 70. Since there are phone changes in Cantonese, some of the consonants are clustered as well. The cross-lingual phones are listed in Table 3.

To include Cantonese accents in English words, the phonetic sequence of the English lexicons in the pronunciation dictionary is modified. The modifications are based on the pronunciation changes mentioned in section 2.1. Some of the lexicons may have different types of variations. Thus, there will be multiple entries for the same lexicons. The dictionary contains an average of 2.267 different pronunciations for each English lexicon.

Table 2. *Acoustic model sets*

Model set	Training Data	Type
A	TIMIT, CUSENT	language dependent
B	CUSENT, CUMIX	language dependent
C	CUSENT, CUMIX	language independent

Table 3 *The cross-lingual phone models*

Phone type	Phone models in IPA
Consonants	f, h, k(kʷ)*, kʰ(kʷʰ)*, l(initial_n)*, m, initial_m, final_m, final_n,ŋ, initial_ŋ, final_ŋ, initial_null, p, pʰ, s, ʃ, ts, tʃ, tsʰ, tʃʰ, t, tʰ, w, j, eng_t, eng_d, eng_k, r, z
Vowel	a, ɐ, ɔ, ɛ, ø, i, ɪ, œ, u, ʊ, y, iu, ai, ɐi,ɐu, ou, ɔi, ei, ʌ, el
Vowel-consonant	ɐp,ɐt,ɐk, ap, at, ak, ɛp, ɛt, ɛk, ut, uk, yt, ip, it, ik, ɔt,ɔk,øt,œk

* clustered phones

4. Language modeling

4.1. Data collection for language modeling

A 6.8M character text database is collected from local newspapers, magazines, newsgroups, and online diaries. The size of the database is small, since there are many constraints on data collection. Standard Chinese is commonly used in the text materials, but it is quite different from spoken Cantonese. Code-mixing text data between standard Chinese and English is therefore not suitable. Mixing between standard Chinese and spoken Cantonese is another problem, since this will involve different sets of lexicons and grammar. Moreover, the frequency of code-mixing is domain specific which mainly occurs in areas that have high interaction with the western culture. There is no list for the commonly used code-switch words; hence, instead of searching for code-mixing text data, we searched for spoken Cantonese text. Articles that contain the selected spoken Cantonese characters (those do not appear in standard Chinese, e.g. 咁,佢,嘅) are selected. Among the collected data, 10% of them are code-mixing.

4.2. Different approaches for language modeling

Although the collected data already covered 8,000 code-switch words, they still cannot include all the existing code-switch words, especially for technical terms, brand names, and people's names. Zero occurrences will occur for the unseen words and will lead to an extremely low language model score. To solve this problem, four language models are proposed which will handle the code-switch words differently. All the language models are tri-gram, which is character based for Cantonese.

- monolingual language model (CAN_LM) – consider all English words as out-of-vocabulary (OOV).
- code-mixing language model (CS_LM) – all English words share the same probability.
- class-based language model (CLASS_LN) – classify the English words into 13 classes according to their part-of-speech (POS) and meaning. The classes are: adjective, companies, date and time, event and activities, fashion, food, brand name, objects and tools, human name, place, sentence and phrase, shops and restaurants, software, verb and the remaining nouns. Most of the classes are nouns since they are in major among code-switch words.
- translation-based language model (TRANS_LN) – translate the English words into their Cantonese equivalent if available; otherwise, use the classes in CLASS_LM. The language model is still character-based, even if the corresponding Cantonese contains multiple characters.

4.3. Evaluation of the language models

Cantonese is homophonic and polyphonic, which means that the mapping between syllables and characters is not one-to-one. Therefore, language models play an important role in selecting the most appropriate characters in the decoding process. The performance of language models is measured by the phonetic-to-text (PTT) conversion rate. This conversion is similar to the second pass of the recognizer. The syllable transcriptions are translated to character transcriptions, and then the hypothesis is compared to the reference. The language model with the minimum PTT error rate will be selected.

5. Language boundary detection

Language boundary detection (LBD) estimates the start and end time of the language segments. In our previous work [13], LBD based on bi-phone likelihood was proposed. The performance of the phone-based LBD system obtains a satisfying result for true code-switch utterances, the code-switch words for which are seldom included in the matrix language. However, when there are accents, the syllable structure of the code-switch words changes. Therefore, the English words would sound like Cantonese words. To tackle problems due to accents, larger units should be considered. Hence, we propose to use a syllable-based LBD, or apply LBD algorithms to the lattice generated by a bilingual speech recognizer.

The syllable-based approach recognizes the input utterance with a syllable recognizer. Bi-syllable likelihoods are calculated from the monolingual Cantonese text database, and the threshold for language identification is derived from development speech data. If the bi-syllable likelihood is larger than the threshold, the syllable pair will be considered as Cantonese; otherwise, it will be considered as English or the language boundary.

The approach based on lattice searches the English word with the longest (W_E) duration from the word lattice. The lattice is generated by the bilingual speech recognizer, which is word based for English and syllable based for Cantonese. The start and end time of the word W_E will be considered as the hypothesis language boundaries of the English segment.

Finally, the hypothesis language boundaries information will be compared. If the errors in both boundaries are smaller than the threshold, they will be regarded as correct LBD.

6. Integration of the sub-systems

The code-mixing speech recognizer is a two-pass system. MFCC features will be extracted from the input waveform, and then they will be passed on to a bilingual speech recognizer, which is syllable based for Cantonese and word based for English. No language models are applied in the first pass. A lattice will be generated by the bilingual speech recognizer, and language boundary information will be integrated to the lattice by re-scoring the acoustic scores of the hypothesis words. Language model scores will finally be integrated to the lattice, and the GWPP will be derived. A character-based hypothesis will then be obtained by best path searching according to the GWPP score.

6.1. Rescoring of the acoustic score

The language boundary information will be compared to the lattice. If the language of the hypothesis in the lattice is identical to the LBD result, a positive weight will be added and hence, the hypothesis will be more likely to be selected in the decoding process. Otherwise, a negative weight will be added, since the hypothesis is in the wrong language. An optimum weight is derived by the development data in CUMIX.

6.2. Generalized Word Posterior Probability (GWPP)

The GWPP re-weights the acoustic and language model likelihood. Acoustic scores are unbounded, while language model scores are limited from 0 to 1 when statistical language models are utilized. In addition, acoustic scores are computed for each frame, while language model scores are computed for each word. Hence, it is necessary to re-weight the two scores so as to obtain a better result.

7. Experimental results

7.1. Acoustic modeling

Three acoustic model sets are proposed, and their performance is compared in Table 4. Monolingual Cantonese data and Cantonese-English code-mixing data from a testing set of CUMIX are employed for evaluation. No language model is applied and the dictionary contains both Cantonese and English.

Table 4 Recognition accuracy of the acoustic model sets

Data	Accuracy	A	B	C
Code-mixing	Can. Syllable and English	57.25%	45.43%	*59.93%*
	Can. Syllable	*60.92%*	45.91%	59.68%
	English	18.86%	40.46%	*58.96%*
Cantonese	Can. Syllable	*63.41%*	56.69%	54.10%

Another experiment is performed on the code-switch words only. Model set C is applied on the same speech recognizer, and the dictionary only includes English words. The recognizer is applied on the English segment of the code-mixing utterances in order to get the upper bound accuracy of the English words. The code-switch words are extracted from the code-mixing utterances which mean that the language boundary information is correct. The word accuracy of this experiment is 81.07%. It shows that language boundary information can greatly improve the recognition accuracy.

7.2. Language modeling

The PTT conversion rate is calculated for the testing data in CUMIX, and the results are listed in Table 5. Unlike conventional speech recognition experience, class-based language model outperforms the others. The main reason comes from the small size of the text database. There are only limited code-mixing text data in the database, hence there may not be adequate data to train the word-based and character-based language models.

Table 5 PTT conversion rate of the language models

Language Model	PTT conversion rates
CAN_LM	88.84%
CS_LM	89.28%
CLASS_LM	*91.52%*
TRANS_LM	86.14%

7.3. Language boundary detection

The language boundary detection accuracy of the three LBD systems is listed in Table 6. Code-mixing testing data from CUMIX are utilized to evaluate the LBD systems.

Table 6 Language boundary detection accuracy

LBD system	LBD accuracy
Phone-based	54.8%
Syllable-based	65.7%
Lattice-based	83.2%

7.4. Character accuracy and English word accuracy

Solutions which obtain the local optimum are selected for the overall system. The cross-lingual acoustic models trained by monolingual and code-mixing speech data will be applied for speech recognition. The language boundary information obtained from the lattice will be based on the output of this speech recognizer in order to search for the English word with the longest duration. The class-based language model will be utilized as well, since the PTT error rate is in the minimum. The summary of the results is listed in Table 7.

With the use of the GWPP, the weight of the acoustic models and the language models is no longer one to one. Instead, the optimum weight of the language models is much higher than that of the acoustic models, since PTT accuracy (Table 5) is much higher than syllable and word accuracy (Table 4).

Table 7 Summary of the experimental results

	Overall accuracy	Character accuracy	English Word accuracy
no LBD	55.29%	56.01%	48.40%
with LBD	56.04%	56.37%	52.99%

8. Conclusion

From the experimental results, it is found that the error mainly comes from the speech recognizer in the first pass. Syllable accuracy and word accuracy are low as compared to the monolingual speech recognizers trained by monolingual data. One of the reasons comes from the speaking style of the speech data. Code-mixing mainly occurs in spontaneous speech, and there will be more phone change and syllable fusion. Phone change, syllable fusion, and Cantonese accents in English words may lead to errors in the phonetic transcriptions of the training and testing speech data. Therefore, the accuracy of the acoustic models is relatively lower for spoken Cantonese and code-mixing speech. More speech data may be necessary in order to study the effect of the accents as well as the speaking style.

It is found that when language boundary information is applied, improvement can be obtained, especially for the code-switch words. It is because the code-switch words only occupy a low percentage of the whole utterances, and they are easily misrecognized as words in the matrix language. The duration of English words is longer than that of Cantonese characters, since Cantonese is monosyllabic. Hence, the lattice-based LBD algorithm obtains a higher LBD accuracy. The LBD information increases the likelihood of the code-switch words to be selected in the decoding process. When the correct language boundary is obtained, the accuracy of the code-switch words can be increased by 28%. Therefore, studies on language boundary detection are necessary for further research.

9. Acknowledgement

This work is partially supported by a RGC grant and a grant awarded by the Shun Hing Institute of Advanced Engineering

10. References

[1] John Gumperz, Discourse Strategies, Cambridge University Press, pp.59, 1982

[2] Peter Auer, Code-Switching in Conversation: Language, Interaction and Identity, Routledge, London, 1998

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

[4] Helena Halmari, Government and Codeswitching: Explaining American Finnish, J. Benjamins, Amsterdam, 1997

[5] Wak-Kit Lo, Frank Soong and Satoshi Nakamura, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels", in Proc. of ISCSLP 2004, pp. 13-16, Hong Kong, 2004

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

[7] Stephen Mattews and Virginia Yip, Cantonese: a Comprehensive Grammar, Routledge, London, 1994

[8] Don Snow, Cantonese a Written Language: the Growth of a Written Chinese Vernacular, Hong Kong University Press, Hong Kong, 2004

[9] Wai Yi Peggy Wong, "Syllable fusion and speech rate in Hong Kong Cantonese", in Proc. of Speech Prosody 2004, pp. 255-258, Nara, Japan, 2004

[10] W. K. Lo, Tan Lee and P. C. Ching, "Development of Cantonese spoken language corpora for speech applications", in Proc. of ISCSLP 1998, pp. 102-107, Singapore, 1998

[11] Joyce Y. C. Chan, P. C. Ching and Tan Lee, "Development of a Cantonese-English Code-mixing Speech Corpus", in Proc. of Eurospeech 2005, pp. 1533-1536, Lisbon, 2005

[12] The CMU Pronuncing Dictionary v0.6, The Carnegia Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

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