Review of “Automatic Recognition of Cantonese-English Code-Mixing Speech”

Shu Ou
University of Washington
sjou@uw.edu

Abstract

This review presents a summary of the study “Automatic Recognition of Cantonese-English Code-Mixing Speech”, detailing issues of the current state-of-the art Automatic Speech Recognition Systems (ASR) in addressing Cantonese-English Code-Mixing Speech; suggested approaches in resolving such issues, and the results of such approaches. In addition, a critical analysis is included in this review to discuss the obstacles the researchers mention, the practicality and potential applications of the researchers’ approaches, as well as recommended follow-up studies that can be done to address some unstated issues of the study.

1 Introduction

The researchers start the study by introducing the concept of code-switching and code mixing, as well as characteristics of the current Cantonese-English Code-Mixing Speech which then leads to a discussion on unaddressed issues of state-of-the art Automatic Speech Recognition (ASR) Systems in processing code-mixing speech.

1.1 Code-Mixing Introduction

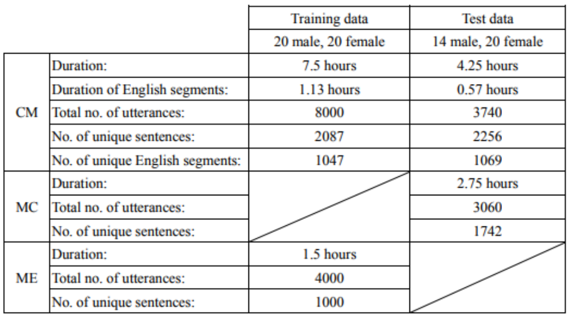
The Code-switching and code mixing concept refer to a type of speech that is commonly found in bilingual speakers in which two or more languages or dialects are alternated within the same speech utterance. The Cantonese-English Code-Mixing speech analyzed in this study refers to a speech habit of English and Cantonese bilingual speakers in Hong Kong in which English terms are often used to better state meanings and emotions when speaking in Cantonese. In this case, Cantonese is the matrix language that provides grammatical frameworks and majority of the lexicons, while English is the secondary/embedded language. The term code-mixing is employed by researchers when describing the Cantonese-English linguistic phenomenon in the study due to the intrasentential nature of this particular speech exchange.

1.2 Issues with ASR Systems in Handling Code-Mixing Speech

The problems of employing the current ASR Systems to analyze Cantonese-English Code-Mixing speech are the following, as described in the paper. Firstly, the current ASR Systems are built based on the assumption that input speech to the System consists of only one language, with identity of such language provided by the user, and that code-mixing handling is rarely considered when developing acoustic and language models for ASR Systems. Secondly, code-mixing speech is not a simple language insertion to other language; rather, it’s an integration of speech where the phonetic properties of one language influences the other. In this case, English words in Cantonese - English Code-Mixing speech are often Cantonese accented, and syllable structures of English word is changed to fit into the Cantonese syllable structures by employing phone insertion or deletion. The conventional large-vocabulary continuous speech recognition (LVCSR) system simply lacks the capability to resolve the issues mentioned.

2 The Approach

In the attempt of addressing the issues mentioned, the researchers propose a modified, two-passed LVCSR system that makes use of a bilingual pronunciation dictionary, cross-lingual acoustic and language models, and syllable lattice based language boundary detector (LBD). The first pass generates a syllable lattice using acoustic models



		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	

Table 6. A Summary of CUMIX.

and pronunciation dictionary, the second pass relies on language models to decode character sequence. Comparisons between two or more approaches in creating the mentioned LVCSR System components are included to ensure the most effective method is used.

2.1 Pronunciation Dictionary

To address the issue in which English phonemes in Cantonese-English Code-Mixing speech commonly adapts features of Cantonese phoneme articulations, pronunciation variations of such English words are taken into consideration when developing the pronunciation dictionary -- Cantonese-English Code-Mixing speech corpus (CUMIX). The CUMIX features three major components: a) Cantonese-English Code-Mixing utterances (CM) based on local newspapers, online newsgroups and diaries; b). Monolingual English utterances (ME) which consists of common English words and phrases that are used in Cantonese-English Code-Mixing speech; and c). Monolingual Cantonese utterances (MC) with CM English segments replaced by the corresponding Cantonese words. The corpus is then divided into two parts with part one being the training data and part one being the testing data. The detailed division is described in Table 6 above.

2.2 Acoustic Modeling

Three acoustic models, ML_A, ML_B, and CL, are established and compared by researchers to find the most suitable model for the proposed LVCSR System. Model ML_A (monolingual A) and Model ML_B (monolingual B) are both language dependent phoneme models, while CL is the cross-lingual acoustic model. The difference between ML_A and ML_B is that ML_A is trained with standard English and Cantonese speech data found in large-scale monolingual speech databases TIMIT and CUSENT, whereas ML_B is trained with English speech data from the CUMIX database and Cantonese speech data from both the CUSENT and CUMIX database. Although CL is trained with the same data as ML_B, the approach is different. CL includes all Cantonese phonemes and only 7 English-specific phonemes, while the other English phonemes are mapped to some Cantonese equivalents, considering Cantonese-Accented English shares features of Cantonese pronunciations. The three acoustic models are evaluated by word recognition experiments using testing data from the CUMIX. Results are described in the Result section.

2.3 Language Modeling

Four languages models are developed in evaluating the best language modeling approach – CAN_LM, CM_LM, CLASS_LM, and TRANS_LM. CAN_LM is a mono-lingual Cantonese language model with English word removals from training text during evaluation; CLASS_LM is a code-mixing language model in which all English words share the same probability; and CLASS_LM is a class-based language model where code switched segments were divided into 15 classes based on their parts of Speech (POS) meanings; lastly, TRANS_LM is translation-based language model that code-switched segments are translated to Cantonese in cases when there are Cantonese equivalents for such segments. Evaluations of language models are done in the Phonetic-to-Text(PTT) conversion task by firstly replacing hypothesized syllable lattice in the first-pass of the LVCSR system with the true syllable-level transcription; in other words, assuming 100% accuracy is achieved by the acoustic models; and then such transcriptions are fed to the presented language models for word sequence determination.

2.4 Language Boundary Detector (LBD)

Language Boundary Detector is crucial in determining the start and end time of individual language segments. The researchers implement two language boundary detectors in this study, and language boundaries generated from these approaches (hypothesized start and end time) are compared to the true language boundaries of the training data. The first LBD is based on syllable bi-gram where bi-syllable likelihoods are calculated from the Cantonese text database, and are used to develop threshold for Cantonese language

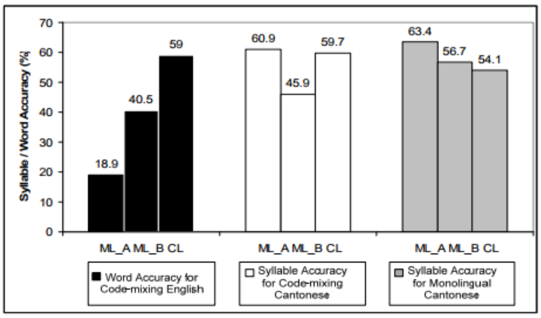


Figure 3. Syllable/word accuracy of the three acoustic models.

identification. A Cantonese syllable recognizer (acoustic models & pronunciation dictionary) is used to retrieve the syllable bigram found in the input. If such bi-syllable probability is larger than the calculated threshold, such syllable pair will be categorized as Cantonese, otherwise, it will be considered as at the language boundary. The second LBD is based on syllable lattice. This approach first employs the bilingual speech recognizer to generate English word lattice, and then such word lattice is searched to find English word with the longest (WE) duration from the word lattice. The start and end time of such duration is considered the language boundaries of the English segment.

2.5 Integration of LVCSR System

Lastly, the proposed LVSCR System is developed by the researchers by integrating the proposed pronunciation dictionaries, cross-lingual acoustic model, cross-lingual language model, Language Boundary detector, and with the use of Generalized Word Posterior Probability to help obtain the most probable sentence hypothesis.

3 Experimental Results

The experimental results are analyzed for different LVSCR System components.

3.1 Acoustic Model

A summary of word recognition accuracies of the three acoustic model is showed in Figure 3. The cross-lingual model has the best overall performance result out of the three proposed acoustic models, in that it correctly recognizes 59% of tested Code-mixing English speech, 59.7% of code-mixing Cantonese and 54.1% of monolingual Cantonese speech utterances. Thus, the cross-lingual model is selected as the acoustic model used in the LVCSR System.

3.2 Language Models

The PTT conversion rates by four proposed language models are within the range of 86% to 91.5%. The similar PPT rate is due to minor differences (with or without the one English code-switched segment) in the transcription data. The Class-based language model is chosen as it has the highest accuracy of 91.5%.

3.3 Language Boundary Detector

The evaluation of the proposed LBDs shows that LBD based on syllable lattice outperforms LBD based on syllable bigram, and achieves an 82% accuracy in determining boundary of Cantonese-English Code-Mixing test utterances. LBD based on syllable bigram only attains a 65.9% accuracy. Thus, LBD is selected to be utilized in the proposed LVCSR System.

3.4 Overall Performance of LVCSR System

The evaluation of the LVCSR System is performed by utilizing the CM and MC test data from CUMIX. The proposed speech recognition systems is able to achieve an a 56.4% accuracy on Cantonese syllables as well as a 53% accuracy on English words in the tested Cantonese-English Code-Mixing speech dataset.

4 Problem Statement and Literature Review Evaluation

Descriptions of current obstacles in processing Cantonese-English Code-Mixing speech are clearly laid-out and are backed-up by real word examples. The mentioned challenges have also been seen later in the experiments when researchers compare various acoustic and language models in the attempt of resolving such issues.

In addition, Prior researches and studies have been reviewed throughout the creation of the LVCSR System for benchmarking proposes as well as to ensure best practice and practicality of research methods. As the field of ASR System on Cantonese-English Code-Mixing speech is fairly new, researchers refer to prior researches done on speech recognition of code-mixing between Mandarin Chinese (matrix language) and other Chinese dialects, and focus their reviews on pure linguistic studies concerning commonality and patterns of Cantonese-English Code-Mixing speech.

When creating the CUMIX data, researchers consult analyses done in previous linguistic studies on Cantonese-English Code-Mixing speech to ensure data in the corpus maintains a realistic percentage distributions of various English word classes (43%, 24%, 13%). Cross-lingual acoustic models developed in previous research to process Mandarin Chinese-Taiwanese Code-Mixing Speech is referenced in the creation of acoustic models for this experiment. Furthermore, when selecting text data to be included in the language modeling, researchers consult prior Cantonese linguistic studies to identify Chinese characters that are frequently used in Cantonese speech but not in standard Chinese, and employ text materials that contain such characters to be used in facilitate statistical language modeling.

5 Experimental Approach and Results Evaluation

The experimental approach is considered convincing and well-organized as the approach implements the framework of LVCSR System, which has widely used in the today's ASR Systems. In addition, the approach of this study has fair experimental conditions in which all acoustic and language models are evaluated based on the same set of test data. Furthermore, methods from this experiment are benchmarked by best practice approaches and discoveries found in prior linguistic studies on Chinese language speech recognition and Cantonese-English Code-Mixing speech.

The experimental results also convincingly demonstrate the benefits of the approach, as comparisons of various acoustic models, language models and LBDs are included in the experiment, results of each model are analyzed to ensure that they are not influenced by unaccounted factors, and models with the highest accuracies are carefully chosen to be integrated into the LVCSR System.

6 Potential Application

Code-mixing and Code switching is a highly dynamic linguistic phenomenon in which there is no "one-size-fits-all" solution available. Characteristics of various language combinations in code-mixing speech can be unique and will need to be carefully analyzed before developing its own LVCSR System. Although the exact approach of the LVCSR System in this study cannot be applied directly to other code-mixing speech, it provides an excellent reference in analyzing code-mixing speech with a Chinese dialect as the matrix language, and English as the secondary language. In the past, majority of the ASR studies were done on code-mixing speech between Chinese dialects, and few studies have been done on code-mixing between Chinese and another typologically distant language. Thus, this study serves as a good starting point for similar studies.

7 Limitations and Future Directions

There are several limitations in the approach employed by the researchers, follow-up studies are recommended to account for those unaddressed factors. One of the limitations is driven by a lack of large database available for Cantonese-English Code-Mixing speech and text data, which has a negative impact on the accuracy of acoustic and language models in the experiment. It's recommended that a database for such speech needs to be built at this point to allow overtime development to ensure more accurate evaluations and recognitions of Cantonese-English Code-Mixing speech. In addition, as mentioned in the paper, Cantonese (also known as the Yue dialects) is a common dialect used in Southern China. Within the Yue dialects family, there are multiple sub-branches such as Guangfu dialects and Sanyi Dialects. Although many of the sub-branches are considered mutual intelligible, there are still differences in certain pronunciation and lexicons used (Wang and Sun, 2015). The study reviewed focuses on addressing Guangfu dialect – English Code mixing speech, and has yet to take into consideration of accent variations of other Cantonese sub-dialects. Thus, a follow up-study that creates a speech corpus which includes data from various sub-branches of Cantonese dialects as well as develops a cross-lingual language model that addresses accented Cantonese-English Code-Mixing speech is recommended to further increase speech recognition accuracy of the current cross-lingual speech recognition models. Lastly, this study concentrates its effort on studying code mixing utterances with only one English segment. As English words are integrated into Cantonese Grammar, there are cases where two or more English segments appear in the same utterance. In particular, such phenomenon often happens in the VERB-NEG-VERB question structure in Cantonese (Wong et al. 2009). For example:

你 ha 唔 happy 啊?
2SG happy NEG happy P.Q.
“Are you happy?”

It’s recommended that further studies are needed to investigate integration of English loan-words into Cantonese Grammar to improve recognition accuracy.

References

Cathy Wong, Robert Bauer, Zoe Lam. 2009. The Integration of English Load-words in Hong Kong Cantonese. *Journal of the Southeast Asian Linguistics Society,* 1:251-266. http://ira.lib.polyu.edu.hk/handle/10397/5824

Joyce Chan, P.C. Ching, Tan Lee, and Houwei Cao. 2019. Automatic Recognition of Cantonese English Code-Mixing Speech. *Computational Linguistics and Chinese language Processing (Volume 14). page 281-304* http://www.aclweb.org/anthology/O/O09/O09-5003.pdf

William Wang, Chaofen Sun. 2015. *The Oxford Handbook of Chinese Linguistics*. Oxford University Press, Oxford, UK.