Effects of Language Mixing for Automatic Recognition of Cantonese-English Code-Mixing Utterances

Houwei Cao, P. C. Ching and Tan Lee

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

Abstract

While automatic speech recognition of either Cantonese or English alone has achieved a great degree of success, recognition of Canton-English code-mixing speech is not as trivial. This paper attempts to analyze the effect of language mixing on recognition performance of code-mixing utterances. By examining the recognition results of Canton-English code-mixing speech, where Canton is the matrix language and English is the embedded language, we noticed that recognition accuracy of the embedded language plays a significant role to the overall performance. In particular, significant performance degradation is found in the matrix language if the embedded words can not be recognized correctly. We also studied the error propagation effect of the embedded English. The results show that the error in embedded English words may propagate to two neighboring Cantonese syllables. Finally, analysis is carried out to determine the influencing factors for recognition performance in embedded English.

Index Terms: code-mixing, automatic speech recognition, error propagation analysis, factor analysis

1. Introduction

Code-switching and code-mixing are common used by people residing in Hong Kong. According to John Gumperz [1], code-switching is "the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems". Different combinations of languages are found in code-switching, such as Spanish-English in the United States, German-Italian and French-Italian in Switzerland, and Hebrew-English in Israel [2]. In Taiwan, code-switching between Chinese dialects, namely, Mandarin and Taiwanese, has become popular in recent years [3]. In Hong Kong, most people, especially the young generation, are Cantonese and English bilinguals. English words are frequently embedded into colloquial 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 [4]. Here, Cantonese is the primary language, also known as the matrix language, while English is the secondary language, also known as the embedded language [5].

In recent years, code-switching/code-mixing has attracted more attention. Many studies are found in the fields of sociolinguistics, psycholinguistics, and general linguistics [4][5][6]. At the same time, speech technologies for processing code-mixing speech have also been developed, such as automatic speech recognition (ASR) [7][8] and text-to-speech synthesis (TTS) [9]. Compared with the monolingual ASR and the conventional multilingual ASR, code-mixing ASR is more challenging because of the unknown language boundaries. From the results reported in [8], there is significant degradation from monolinguall ASR to code-mixing ASR. To improve system performance, many studies have been carried out on the design of acoustic models, lexical and language models, as well as decoding algorithms. In contrast, few studies have been carried out to further investigate the code-mixing effect in speech recognition in a systematic way. In this paper, analysis have been done to further investigate how and why the degradation caused specifically by code-mixing. The analysis is mainly based on the speech recognition results.

The organization of this paper is as follows. After a short review of the phonological structure of Cantonese and English in Section 2, the speech recognition experiments are described in Section 3. In Section 4, the recognition results are reported, and the effect of embedded English is discussed. An analysis of error propagation in embedded English is performed as well. The Factor analysis is carried out in section 5 to determine the influencing factors for recognition performance in code-mixing English. Finally, some implications for better code-mixing ASR are derived in conclusion.

2. Phonological Structure of Cantonese and English

Cantonese is a Chinese dialect, mainly spoken in southern China, Hong Kong, and Macau. It is a monosyllabic and tonal language. The general syllable structure of Cantonese is C1-V-C2, where C1 and C2 are consonants, and V is either a vowel or a diphthong. Since C1 and C2 are optional, all Cantonese syllables take the forms V, C-V, C-V-C or V-C.

Meanwhile, English is an Indo-European language. The composition of syllables in English has greater variations than Cantonese. About 80% of the syllables in English take the C-V-C structure, and the remaining 20% could be CCV, VCC, CCCV, CCCVCC, and so on.

3. Speech Recognition Experiments

3.1. Speech Corpora

The speech materials used in this study come from two databases. They are the monolingual Cantonese corpus CUSENT [10] and the Cantonese-English code-mixing corpus CUMIX [11]. CUSENT is a read-style speech corpus of Cantonese. It is intended for the development of speaker-independent continuous speech recognition for Cantonese. Sixty-eight native Cantonese speakers are involved, and the amount of speech data is 20 hours. CUMIX is a database developed specifically for code-mixing speech recognition. The
spoken contents in CUMIX are mainly daily conversation or jargons by university students in Hong Kong. There are three different types of utterances in CUMIX: (1) Cantonese-English code-mixing utterances (CM), (2) Monolingual colloquial Cantonese utterances (MC), and (3) Monolingual English words and phrases (ME). It contains 16 hours of speech data from 74 speakers. The training data include utterances from 20 male and 20 female speakers. Each speaker has 200 CM utterances and 100 ME utterances. There are 14 male and 20 female speakers in test data. Each of them has 120 CM utterances and 90 MC utterances. The unique code-switched words in this corpus is 1047 in the training data, and 1069 in the test data.

3.2. Baseline Speech Recognizer

A set of cross-lingual acoustic models is designed. We use IPA to facilitate an intuitive comparison between Cantonese and English phonemes. Some of the phonemes in Cantonese and English are labeled with the same IPA symbols by phoneticians. These phonemes are expected to be phonetically very close to one another and they are merged into the same phone classes in our experiments. A total of 65 phonemes are selected for the acoustic models. 17 phonemes are shared between two languages. 22 are English-specific and the remaining 26 are Cantonese specific.

The acoustic models are tri-phone HMMs trained by 29 hours of training speech from the CUSENT and CUMIX databases. The acoustic feature vector consists of 12 MFCC coefficients, log energy, and their first and second-order derivatives. Each state in HMM has 16 Gaussian mixtures.

In our application, the mother tongue of the speaker is Cantonese. It is inevitable that the embedded English words carry Cantonese accent to certain extent. In many cases, the syllable structure of an English word is modified to become acceptable in Cantonese, this may be achieved by insertion or deletion of phones. In addition, there also exist phone changes in English-accented English. To reflect the effect of Cantonese accent in English words, the pronunciation dictionary is modified based on our previous study on pronunciation variation [12]. The dictionary contains an average of 1.48 pronunciations for each English lexical item.

4. Speech Recognition Results and Analysis

4.1. Code-mixing vs. Monolingual

The recognition accuracy is given in Table 1. Cantonese-English code-mixing and monolingual colloquial Cantonese utterances from the test set of CUMIX are employed for evaluation. The recognition performance is compared with syllable accuracy for Cantonese and word accuracy for English. No language model is applied, and the lexicon contains both Cantonese and English entries. Another experiment is performed on embedded English words only. It attains a recognition accuracy of 76%. This is the upper bound accuracy with perfect language boundary information.

Table 1: Baseline recognition accuracy on code-mixing and monolingual speech data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>57.3%</td>
<td>57.6%</td>
<td>54.1%</td>
</tr>
<tr>
<td>Monolingual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cantonese syl. Acc.</td>
<td>62.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English word Acc.</td>
<td></td>
<td>76%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Evaluation results in terms of the recognition accuracy on embedded English

<table>
<thead>
<tr>
<th></th>
<th>Set A</th>
<th>Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly rec. Eng.</td>
<td>66.5%</td>
<td>62.4%</td>
</tr>
<tr>
<td>Wrongly rec. Eng.</td>
<td>46.7%</td>
<td>59.2%</td>
</tr>
<tr>
<td>Oracle experiments</td>
<td>100%</td>
<td>94.5%</td>
</tr>
</tbody>
</table>

It is noticed that the recognition performance is quiet different between Set A and Set B. If the embedded English words can be recognized, the recognition accuracy on code-mixing Cantonese is 63.2%, which is comparable to the monolingual results shown in Table 1. This means that the embedding effect would not degrade the performance of code-mixing ASR if the embedded segments can be recognized correctly. On the other hand, if the embedded words can not be recognized, the error on embedded language will lead to significant degradation on the matrix language.

The recognition accuracy attained by the oracle experiment is 59.2% and 94.5% for code-mixing Cantonese and English, respectively. Recognition errors still exist on code-mixing English since the oracle experiment does not specify the language boundaries. The assimilation caused by code-mixing makes some short English words very similar to Cantonese syllables, and therefore, some of the Cantonese syllables are wrongly decoded as the reference English, while the real English segments are recognized as Cantonese. For code-mixing Cantonese, the performance attained by the oracle experiment shows obvious improvement, although it is still lower than that of the monolingual case. This shows that the information about the embedded language is very useful.
4.3. Error Propagation of Embedded English

Analysis is carried out to investigate the error propagation of the embedded English.

We divide the recognition results on code-mixing utterances into three cases in terms of the position of the embedded English segments. For each case, the error propagation observed under the following conditions: (1) when English was recognized correctly (Set A), (2) when English was wrongly recognized (Set B), and (3) the oracle experiment for Set B. The error propagation graphs are given in Figure 2. Graphs (a), (b) and (c) show the error propagation when the embedded English words are in the beginning, middle, and end of the code-mixing utterances, respectively.

It can be observed that the embedding effect would not induce error nearby if the embedded segments can be recognized correctly. However, if the English words were wrongly recognized, the error would propagate to the surrounding Cantonese syllables. In particular, when the English words appeared in the beginning/end of the utterances, the error would significantly propagate to the first and second following/preceding Cantonese syllables, respectively. When the English words are in the middle of the utterances, the errors would slightly spread to the first preceding Cantonese syllable, but this would seriously affect the first subsequent Cantonese syllable. In oracle experiment, embedded English will only propagate at most one neighboring Cantonese syllable.

Another analysis is performed to reveal the degree of the embedded language effect with different error types in English. In this study, we classified the error on code-mixing English into three types as follows.

- Error type A: the whole English segment is recognized as another English segment.
- Error type B: the whole English segment is recognized as Cantonese syllables.
- Error type C: part of the English segment is recognized as another English segment, the remaining part is recognized as Cantonese syllables.

The recognition accuracy in code-mixing Cantonese with three different English error types is shown in Table 3. It noticed that, only 10% of the errors in English are language-specific errors. In this case, the language boundary information is expected to be correct, and therefore the recognition accuracy in Cantonese would not be affected by the embedded language. On the other hand, 90% of the errors in English are caused by confusion between languages. In this case, additional deletion and insertion error would be introduced by imperfect language boundary information, and therefore, error on the embedded language will propagate.

Table 3: Cantonese accuracy with different English error types

<table>
<thead>
<tr>
<th>English error type</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>10%</td>
<td>49%</td>
<td>41%</td>
</tr>
<tr>
<td>Cantonese syl. Acc.</td>
<td>61%</td>
<td>51%</td>
<td>48%</td>
</tr>
</tbody>
</table>

5. Discussion

Factor analysis is carried out to determine the influencing factors for recognition performance in code-mixing English. Studies are performed using select factors as follows.

- Position of English words
- Phonology of English words
- Characteristics of nearby Cantonese syllables

The recognition performances with different positions are given in Table 4. It is shown that the English accuracy is relatively lower when English is in the middle. This may be due to the bidirectional articulation from the preceding and posterior Cantonese syllables.

Table 4: Recognition results with different English positions

<table>
<thead>
<tr>
<th>Position</th>
<th>Beginning</th>
<th>Middle</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Acc.</td>
<td>56.7%</td>
<td>57.3%</td>
<td>57.2%</td>
</tr>
<tr>
<td>Cantonese syl. Acc.</td>
<td>56.7%</td>
<td>57.6%</td>
<td>57.1%</td>
</tr>
<tr>
<td>English word Acc.</td>
<td>56.3%</td>
<td>53.5%</td>
<td>57.6%</td>
</tr>
</tbody>
</table>

Further analysis is performed to investigate the effects of English phonology. The study is performed in terms of syllable numbers and syllabic structure of English words. Table 5 shows the English accuracy with different syllable numbers per word. It indicates that the recognition accuracy increase with the increase of syllables.

Next, analysis is performed on the syllable structure in terms of number of syllables per English word. We observed...
that if the embedded English is a monosyllabic word, the syllable structure indicates a significant effect on recognition accuracy. The English accuracy is found to be much lower when the C-V-C structure is retained. However, with the increase in syllable numbers, the effect of the syllable structure decreases. If the English word consist of three or more syllables, the recognition performance would not be affected by its syllable structure.

Analysis is carried out to investigate the effect of the immediately preceding/following Cantonese syllables around the embedded English. The study is performed in terms of analysis syllable correctness and the syllabic structure. It is found that there is indeed a strong correlation between English word accuracy and the correctness of the neighboring Cantonese syllables. Table 6 gives the details when English is embedded in the middle of utterances. It shows that the recognition accuracy on English words can achieve over 80% when the surrounding Cantonese syllables can be recognized. However, the recognition performance is degraded to less than 40% if the surrounding Cantonese syllables cannot be recognized. On the other hand, no obvious observation is made in relation to the Cantonese syllable structure.

Table 5: English accuracy with the number of syllables per English word

<table>
<thead>
<tr>
<th>No. of Syllables</th>
<th>Proportion</th>
<th>English word Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.2%</td>
<td>37.4%</td>
</tr>
<tr>
<td>2</td>
<td>45.6%</td>
<td>53.8%</td>
</tr>
<tr>
<td>≥3</td>
<td>32.2%</td>
<td>64.2%</td>
</tr>
</tbody>
</table>

Table 6: English accuracy versus surrounding Cantonese correctness

<table>
<thead>
<tr>
<th>Cantonese</th>
<th>English word Acc.</th>
<th>Cantonese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corr.</td>
<td>82.1%</td>
<td>Corr.</td>
</tr>
<tr>
<td>Corr.</td>
<td>41.8%</td>
<td>Wrong</td>
</tr>
<tr>
<td>Wrong</td>
<td>56.7%</td>
<td>Corr.</td>
</tr>
<tr>
<td>Wrong</td>
<td>37.0%</td>
<td>Wrong</td>
</tr>
</tbody>
</table>

6. Conclusions

This study mainly investigated the effect of embedded language on code-mixing speech recognition in Cantonese-English code-mixing ASR. As discussed in Section 4, the recognition accuracy in code-mixing is significantly affected by the recognition performance of the embedded language. In particular, the embedding effect would not degrade the performance of code-mixing ASR if the embedded segments can be recognized correctly, but significant degradation is found in the matrix language if the embedded words can not be recognized. This indicates that the recognition performance on embedded language is very important, and it is believed that the improvement in the embedded language will bring improvement to the matrix language as well. Future studies to improve the performance of embedded English are therefore desirable.

The analysis of error propagation shows that the error in embedded English words may propagate to two Cantonese syllables nearby. This indicates that the acoustic score on the first two Cantonese syllables nearby the embedded English is unconvincing as compared to that of other Cantonese syllables. This suggests that the acoustic weighting on these two syllables can be decreased, and the other cues such as language information should be emphasized in contrast.

Language boundary information is found to be very useful. Language-specific error in English with correct language boundary information would not affect the nearby Cantonese. Therefore, language boundary information can be used as the confidence score during decoding in future studies.

Factor analysis is carried out to determine influencing factors for recognition performance in code-mixing English. The findings related to the phonology of English can be used to guide better speech recognition.

7. Acknowledgement

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

8. References