Language Modeling for Speech Recognition of Spoken Cantonese

Yu Ting Yeung, Houwei Cao, N.H. Zheng, Tan Lee and P. C. Ching

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

Abstract
This paper addresses the problem of language modeling for LVCSR of Cantonese spoken in daily communication. As a spoken dialect, Cantonese is not used in written documents and published materials. Thus it is difficult to collect sufficient amount of written Cantonese text data for the training of statistical language models. We propose to solve this problem by translating standard Chinese text, which is much easier to find, into written Cantonese. A rule-based method of translation is devised and implemented. Three different language models are trained from different types of text. They are evaluated in the task of LVCSR. Experimental results confirm that the translated text can well represent Cantonese spoken in formal occasions like broadcast news. For colloquial Cantonese, language model adaptation with a limited amount of colloquial Cantonese text data would be a practically feasible solution that leads to reasonable speech recognition performance.

Index Terms: language model, speech recognition, Cantonese

1. Introduction
Lexicon and language model are the key components of a large-vocabulary continuous speech recognition (LVCSR) system. Lexicon specifies the vocabulary items that the system can recognize and their pronunciations. Language model captures the regularities of the language and provides linguistic constraints on the speech recognition output [1]. Lexicon is often derived from dictionaries. Statistical language models are trained with large amounts of text data, which are usually collected from electronic versions of published materials, e.g., newspaper texts [2]. There are assumptions that the lexicon contains all possible words in the speech and the training text data are reliable representations of speech, i.e., they coincide with the true transcription of speech. These assumptions are not trivial. While the text on most published materials are able to represent formal speech, they can be quite different from conversational speech. This is especially true for Chinese dialects.

Mandarin is the official spoken language in mainland China and Taiwan. Standard Chinese is the written form of Mandarin. The text in newspapers are very close to Mandarin, and thus are suitable for language modeling in Mandarin LVCSR systems. On the other hand, many other Chinese dialects neither have standard written forms nor agree with standard Chinese when they are written down directly. Among these dialects, Cantonese is commonly used in southern China, Hong Kong, Macau and overseas Chinese communities. Cantonese speech, when being written down, shows substantial differences from standard Chinese. Text of standard Chinese is not adequate to reflect the linguistic properties of Cantonese. For Cantonese LVCSR, a large amount of “written Cantonese” text data are desirable for language modeling, but they are generally unavailable.

Our research team at the Chinese University of Hong Kong has been working on Cantonese speech recognition since late 1990s. We have created large-scale Cantonese speech databases: CUCorpora [3] and developed an LVCSR system [4] based on these databases. The major limitation of the system is that they deal with only Cantonese read speech with presumably standard Chinese content.

This paper describes our recent efforts towards speech recognition of spoken Cantonese. In particular, we address the problems of language modeling for Cantonese spoken in daily communication. In the next section, we define and clarify the terms “standard Chinese” and “written Cantonese”, as well as various types of Cantonese speech. In Section 3, we describe several language models of Cantonese trained with different types of Chinese text. A rule-based method is adopted to translate standard Chinese text to written Cantonese. Speech recognition experiments are carried out to evaluate the performance of the language models in an LVCSR task and the results are presented in Section 4.

2. Cantonese vs. Standard Chinese
Snow [5] suggested that a fairly accurate description of “written Cantonese” would be “Cantonese speech written out in Chinese characters.” In Hong Kong, written Cantonese is not accepted in official documents. They are not taught at schools. Although the amount of published work using written Cantonese is increasing recent years, text data obtained from these sources are sparse and usually domain-specific, making it practically infeasible to develop language models for truly Cantonese speech.

Despite the diversity of spoken dialects, standard Chinese is the formal written language in the whole Chinese society, including Hong Kong. Standard Chinese is dominant in massive publishing media in Hong Kong, although Snow [5] noticed that standard Chinese in Hong Kong is somehow different from those in mainland China and Taiwan. For example, there are lexical differences like “巴士/baa siр” in Hong Kong and “公交/gong jiao cheр” in mainland, both words meaning “bus” in English. As these types of localized lexicons are commonly used in publishing media, they do not create big problems in language modeling for Cantonese.

There is consensus that some daily used Cantonese words, which are not found in standard Chinese, are not allowed in formal Chinese writing. Table 1 shows some of these examples. We refer them as “Cantonese terms” throughout this paper.

It is noticed that the content of Cantonese depends greatly on the occasion of speaking. In formal occasions, the transcription of Cantonese speech would resemble standard Chinese in vocabulary choices. In more casual situations, Cantonese terms are used frequently. When such kind of Cantonese speech is written down, it could be unintelligible to Mandarin speakers [6]. In this paper, we divide Cantonese speech into two major
### 3. Language Modeling for Cantonese

#### 3.1. Training of different language models

In this study, three sets of tri-gram language models are developed and evaluated. They are trained from two distinct sets of text data. The first set is considered as standard Chinese text. It was obtained from local newspapers, and contains 98-million Chinese characters. About 78% of the content are local, international, financial and sports news. It is used to train the language model for standard Chinese, which is referred to as StdChi. A lexicon with 14,000 words is used for StdChi. These include some characters that are specific to written Cantonese. A language model for formal Cantonese, denoted by FmlCan, is trained by translating standard Chinese text into written Cantonese. The translation is based on a rule-based method. The rules are learnt automatically from a small parallel corpus of standard Chinese and written Cantonese. Details of the implementation will be discussed in Section 3.2.

The second set of text data are regarded as colloquial Cantonese. They were collected from entertainment news, local magazines, newsgroups and online diaries on the Internet. Cantonese terms are frequent in these texts. However, this set of text data contains only 6-million Chinese characters. They are not adequate to train a full language model by themselves. Instead, they are used as adaptation data to modify FmlCan by linear interpolation. The resulted language model is referred to as CollCan. Here is a summary of three language models in Table 2.

<table>
<thead>
<tr>
<th>Name of language model</th>
<th>Lexicon size</th>
<th>Text data for training</th>
<th>Percentage of Cantonese terms in training text</th>
</tr>
</thead>
<tbody>
<tr>
<td>StdChi</td>
<td>14K</td>
<td>standard Chinese text (98M characters)</td>
<td>~0.1%</td>
</tr>
<tr>
<td>FmlCan</td>
<td>9K</td>
<td>written Cantonese translated from standard Chinese text</td>
<td>7.8%</td>
</tr>
<tr>
<td>CollCan</td>
<td>9K</td>
<td>colloquial Cantonese text (6M characters) for adaptation of the FmlCan model</td>
<td>16.8%</td>
</tr>
</tbody>
</table>

#### 3.2. Rule-based translation from standard Chinese text to written Cantonese

A rule-based method is used to translate standard Chinese text to written Cantonese in order to provide training data for the FmlCan model. A parallel corpus of standard Chinese and written Cantonese was developed by manual translation. It started with 8,300 complete sentences (90,000 characters) of standard Chinese (newspaper text). A few university students were asked to make necessary modification on these sentences so that they can be read out as natural Cantonese, with their meanings unchanged. The translated text is about the same amount as the original one. Word segmentation was performed automatically for the parallel sentences. Due to the small size of the corpus, the words in our lexicon were clustered into 10 classes using the Word Exchange Algorithm [8]. From word-level alignment of the parallel sentences, conversion rules are established by

---

**Table 1**: Examples on lexicon differences between Cantonese and standard Chinese.

<table>
<thead>
<tr>
<th>English</th>
<th>Cantonese</th>
<th>Standard Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>(possessive marker)</td>
<td>呢/ni\ entrepreneur \n (perfective aspect marker)</td>
<td>咋/za\ entrepreneur</td>
</tr>
<tr>
<td>(pluralizer for pronouns)</td>
<td>佢/te\ entrepreneur</td>
<td>佢/te\ entrepreneur</td>
</tr>
<tr>
<td>This/these</td>
<td>喺/keoi\ entrepreneur</td>
<td>佢/te\ entrepreneur</td>
</tr>
<tr>
<td>He/she</td>
<td>佢/keoi\ entrepreneur</td>
<td>佢/te\ entrepreneur</td>
</tr>
<tr>
<td>Yesterday</td>
<td>昨日/cam yat\ entrepreneur</td>
<td>昨天/zuotian\ entrepreneur</td>
</tr>
</tbody>
</table>

---

**Figure 1**: The differences between formal Cantonese, colloquial Cantonese and standard Chinese, for the same meaning.
There is a further improvement for CUMIX when used, indicating the effectiveness of language model adaptation (CUMIX), cast news using formal Cantonese. For colloquial Cantonese has the lowest perplexity on HKBN, which contains broad-

Table 3: Character perplexities of the three language models.

<table>
<thead>
<tr>
<th>Language models</th>
<th>CUSENT (Standard Chinese)</th>
<th>HKBN (Formal Cantonese)</th>
<th>CUMIX (Colloquial Cantonese)</th>
</tr>
</thead>
<tbody>
<tr>
<td>StdChi</td>
<td>133.9</td>
<td>261.5</td>
<td>1381.1</td>
</tr>
<tr>
<td>FmlCan</td>
<td>155.8</td>
<td>89.9</td>
<td>384.4</td>
</tr>
<tr>
<td>CollCan</td>
<td>232.9</td>
<td>125.6</td>
<td>127.7</td>
</tr>
</tbody>
</table>

3.3. Perplexity of language models

Perplexity is an important performance index for statistical language models. A reduction in perplexity generally results in improvement on speech recognition performances. For Chinese languages, character perplexity [12] is often used since word segmentation of Chinese sentences tends to be ambiguous. Table 3 shows that character perplexities of the three language models described above. The test data include standard Chinese, formal Cantonese and colloquial Cantonese (see Section 4.1 for more details). The character perplexities suggest the most appropriate language model for each type of text data. For standard Chinese (CUSENT), the perplexity of StdChi is the lowest among the three language models. The FmlCan model has the lowest perplexity on HKBN, which contains broadcast news using formal Cantonese. For colloquial Cantonese (CUMIX), FmlCan has a much lower perplexity than StdChi. There is a further improvement for CUMIX when CollCan is used, indicating the effectiveness of language model adaptation with colloquial Cantonese text data.

4. Recognition Experiments

4.1. Speech corpus

Test speech used in our experiments are from three Cantonese databases. CUSENT is a read-style speech corpus of Cantonese [3]. The speakers were asked to read read prompted newspaper text. It is used to represent Cantonese speech with standard Chinese content. HKBN is a database of Cantonese broadcast news. It contains recordings of evening news reports of Commercial Radio Hong Kong (CRHK) from September 1998 to June 1999. It is used to represent formal Cantonese. CUMIX is a database containing colloquial Cantonese [10]. The spoken contents are mainly daily conversation or jargons by universities students in Hong Kong. A summary of the three corpora is given in Table 4.

Table 4: Three speech corpora used in our experiments.

<table>
<thead>
<tr>
<th>Corpus name</th>
<th>Amount of test speech (in hours)</th>
<th>Number of speakers</th>
<th>Percentage of Cantonese terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSENT</td>
<td>1.10</td>
<td>12</td>
<td>~&lt;0.0%</td>
</tr>
<tr>
<td>HKBN</td>
<td>1.30</td>
<td>14</td>
<td>8.1%</td>
</tr>
<tr>
<td>CUMIX</td>
<td>1.25</td>
<td>14</td>
<td>25.6%</td>
</tr>
</tbody>
</table>

4.2. Speech recognition setup

The language models are evaluated in an LVCSR system. The acoustic models are tri-phone hidden Markov models, trained by 20 hours of training data from the CUSENT database [11]. The acoustic feature vector consists of 12 MFCC coefficients, log energy, and their first and second-order derivatives. Each state in the HMM has 16 Gaussian mixtures. The training data is phonetically balanced [3]. When the test data from CUSENT itself is used, the phone recognition accuracy is about 85.6%.

The decoding for LVCSR has two passes. The first pass generates a word lattice by Viterbi decoding with the tri-phone acoustic models and bi-gram language models. In the second pass, the word lattice is re-scored with tri-gram language models to produce the best word sequence output.

4.3. Recognition accuracy

Table 5 shows the recognition performance attained with the three language models. The performance is in terms of character accuracy. It is clear that the language model trained from the best-matched text data gives the best performance. StdChi for CUSENT, FmlCan for HKBN and CollCan for CUMIX. For CUSENT, the accuracy drops by 4.4% when StdChi is replaced by FmlCan. Whereas for HKBN, StdChi is less appropriate
than FmlCan. This confirms that Cantonese, even if it is spoken in formal occasions, is indeed very different from standard Chinese. This discrepancy, however, is very mild when colloquial Cantonese speech in CUMIX is considered. For CUMIX, both StdChi and FmlCan perform poorly. The recognition accuracy is improved significantly when CollCan is used. As shown in Table 2, the percentages of Cantonese terms in colloquial Cantonese and formal Cantonese are 16.8% and 7.8% respectively. Note that CollCan is not fully trained by colloquial Cantonese text. It seems that language model adaptation with a small amount of data is a feasible way to deal with the data sparseness problem for colloquial Cantonese.

Further analysis of recognition performance is done by separating Cantonese terms from standard Chinese words, as shown in Table 6. It is shown that the improvement of overall recognition performance for test speech of HKBN and CUMIX are largely due to the improved recognition on Cantonese terms, when formal or colloquial Cantonese language models are used. Many Cantonese terms are function words which connect standard Chinese words in an utterance. It is noted that, for colloquial Cantonese speech, recognition with better-matched language models not only improves the accuracy of Cantonese terms, but also the accuracy of standard Chinese words.

Table 5: Character accuracies attained with different language models. The numbers in boldface are those attained with the best match between training text and test speech.

<table>
<thead>
<tr>
<th>Language model</th>
<th>CUSENT</th>
<th>HKBN</th>
<th>CUMIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>StdChi</td>
<td>82.2%</td>
<td>68.8%</td>
<td>50.0%</td>
</tr>
<tr>
<td>FmlCan</td>
<td>77.8%</td>
<td>73.0%</td>
<td>56.0%</td>
</tr>
<tr>
<td>CollCan</td>
<td>71.2%</td>
<td>71.1%</td>
<td>70.4%</td>
</tr>
</tbody>
</table>

Table 6: Character accuracies (%) for Cantonese terms (CC) and standard Chinese words (SC).

<table>
<thead>
<tr>
<th>Language Models</th>
<th>CUSENT</th>
<th>HKBN</th>
<th>CUMIX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>SC</td>
<td>CC</td>
</tr>
<tr>
<td>StdChi</td>
<td>-</td>
<td>82.2%</td>
<td>13.6%</td>
</tr>
<tr>
<td>FmlCan</td>
<td>-</td>
<td>77.8%</td>
<td>51.3%</td>
</tr>
<tr>
<td>CollCan</td>
<td>-</td>
<td>71.2%</td>
<td>57.9%</td>
</tr>
</tbody>
</table>

5. Conclusions

For Cantonese spoken in daily communication, a language model trained from standard Chinese texts is inadequate and written Cantonese texts must be used for an LVCSR system to attain reasonable performance. In the absence of sufficient amount of written Cantonese data, simple rule-based translation from standard Chinese text is proven to be an effective solution. The translated text represents formal Cantonese as those used in broadcast news. To model colloquial Cantonese, we can collect a limited amount of colloquial Cantonese text data and use them to adapt the formal Cantonese language models. The more adaptation data, the better the recognition performance is expected. Our experimental results also show that language models trained from content-matched text improve the recognition of not only the Cantonese terms but also standard Chinese words in the input utterances.

6. Acknowledgements

This work is supported by an Earmarked Grant from the Hong Kong Research Grants Council (Ref.: CUHK 411306) and a project grant from the Shun Hing Institute of Advanced Engineering, the Chinese University of Hong Kong. The authors thank Dr. Brian Mak for providing the CRHK broadcast news database. This database is intended for research only, fostering collaborative work among speech researchers in Hong Kong.

7. References