Deploying Google Search by Voice in Cantonese

Yun-Hsuan Sung, Martin Jansche, Pedro J. Moreno

Google Inc., USA

yhsung, mjansche, pedro@google.com

Abstract

We describe our efforts in deploying Google search by voice for Cantonese, a southern Chinese dialect widely spoken in and around Hong Kong and Guangzhou. We collected audio data from local Cantonese speakers in Hong Kong and Guangzhou by using our DataHound smartphone application. This data was used to create appropriate acoustic models. Language models were trained on anonymized query logs from Google Web Search for Hong Kong. Because users in Hong Kong frequently mix English and Cantonese in their queries, we designed our system from the ground up to handle both languages. We report on experiments with different techniques for mapping the phoneme inventories for both languages into a common space. Based on extensive experiments we report word error rates and web scores for both Hong Kong and Guangzhou data. Cantonese Google search by voice was launched in December 2010.
Index Terms: voice search, Cantonese speech recognition, multilingual speech recognition

1. Introduction

Cantonese is one of the major Chinese dialects spoken by tens of millions of people in Hong Kong, Southern China as well as many overseas Chinese communities. In Hong Kong it is written using traditional Chinese characters similar to those used in Taiwan. Chinese script is much harder to type than the Latin alphabet, especially on mobile devices with small or virtual keyboards. People in Hong Kong typically use the "Cangjie" (倉頡) or "Handwriting" (手寫輸入) input methods. Cangjie has a steep learning curve and requires users to break characters down into sequences of graphical components. The Handwriting method is easier to learn but slow to use. Neither is an ideal input method for people in Hong Kong trying to use Google Search on their mobile phones.

Speaking is generally much faster and more natural than typing. Moreover, some Chinese characters – like 滘 in 滘西州 (Kau Sai Chau) and 咗 in 咗典乍街 (Pottinger Street) – are so rarely used that people often know only their pronunciation, and not how to write them.

Our Cantonese Voice Search begins to address these situations by allowing Hong Kong users to speak search queries instead of entering Chinese characters on mobile devices. We believe our development of Cantonese Voice Search is a step towards solving the text input challenge for devices with small or virtual keyboards for users in Hong Kong.

Large vocabulary Cantonese speech recognition has been studied for a while in the speech research community [1] but to our knowledge no commercial Cantonese voice search system has ever been built. In this paper we describe the strategies and challenges in developing such a system. Some challenges we faced are unique to Cantonese, others typical of Asian languages, and some universal to all languages.

We describe our data collection efforts in Section 2 and provide a system overview in Section 3. Multilingual issues are discussed in Section 4. Finally, we report our experimental results in Section 5.

2. Data Collection

In contrast to English and other resource-rich languages, there are few existing Cantonese datasets that can be used to train a speech recognition system. It is true that some corpora have been collected for building large vocabulary Cantonese speech recognizers [2]. However, these datasets were not specifically developed for mobile voice search applications and are likely to be mismatched in several ways. For example, our acoustic data requirements are quite specific. Our voice search application assumes that audio is spoken into a mobile phone, with multiple types of background noise, etc. The semantic nature of the utterance is also specific to the task. A voice search system expects short spoken search queries, not long dictation or broadcast-news-style sentences. For all these reasons we decided early on that we needed to collect our own acoustic data. The efficient collection of high quality data thus became a crucial issue in system development. In this section we discuss how we collected text and audio data.

2.1. Text Data

In Google Search by voice, users speak their search queries and the system returns search results. Because of the similarities with regular, typed search, one of the best sources of training data for language models are typed search queries. Because the semantic content of search queries varies strongly by region, we based our textual data on anonymized search query logs from our www.google.com.hk search engine.

A sufficient amount of anonymized search query logs are extracted and further processed. One of the first steps of this processing is the normalization of Chinese scripts. Search queries from www.google.com.hk are found in both Traditional and Simplified Chinese script. Traditional Chinese script is preferred in Hong Kong. To support Cantonese for Hong Kong, we first convert queries in Simplified script to Traditional script. At this point we may be tempted to simply filter away queries that are not in Chinese script. However, Hong Kong is a multilingual society and most residents speak both English and Cantonese. A voice search system in Hong Kong that only supports Cantonese would be impractical for users; conversely, a system that only handles English would leave large parts of the population unhappy. Clearly our Cantonese system needs to recognize both Cantonese and English queries. For this reason we also retain queries containing frequent English words, URLs, and other tokens in Latin alphabet.

Table 1 shows the language distribution of search queries in Mainland China, Taiwan, and Hong Kong. Compared to Main-

	Chinese	English	URL	Number
China	84.7	11.6	1.2	2.5
Taiwan	77.4	20.6	0.6	1.4
Hong Kong	60.8	34.8	2.9	1.5

Table 1: Language distribution (%) of search queries in Mainland China, Taiwan, and Hong Kong.

land China and Taiwan, Hong Kong has much higher percentages of English queries (34.8%) and queries with URLs (2.9%).

2.2. Audio Data

In general our voice search system is mainly used on mobile phones in a variety of acoustic environments, including use at home, on the subway, in shopping centers, etc. For optimal performance it is important that data collection follows actual usage conditions closely. It is also important that the demographic distribution of prospective users is well represented during data collection. Finally it is helpful if the spoken utterances we collect are based on popular search queries in Hong Kong.

To address all of these desiderata, we collected spoken utterance in different acoustic environments from a variety of speakers using our DataHound Android application [3], which displays prompts based on common Chinese and English search queries on a mobile device, asks users to speak them back, and records their utterances. We recruited local Cantonese-speaking volunteers to record more than half a million spoken queries. The volunteers included both female and male speakers from different age ranges. We recorded the audio samples in both quiet and noisy environments, including offices, shopping centers, public transportation and others. To increase our Cantonese accent coverage we collected audio data not only in Hong Kong but also in Guangzhou (廣州), the largest city in adjacent Cantonese-speaking Guangdong province in southern China.

3. System Description

3.1. Acoustic Model

Our acoustic models are standard 3-state context dependent (triphone) models with a variable number of Gaussians per state. These are trained on a 39-dimensional vector composed of PLP cepstral coefficients and their first and second order derivatives. Cepstral mean normalization is applied as well as an energy based endpointer to remove excessive silence. Our frontend also uses Linear Discriminant Analysis (LDA). We use standard decision-tree state-based clustering followed by semi-tied covariance (STC) modeling [4] and an FST-based search [5]. Our acoustic models are gender independent, maximum-likelihood trained, followed by boosted MMI (BMMI) [6].

Cantonese, like closely-related Mandarin, is a tonal language. The inventory of tones is much larger in Cantonese than in Mandarin, which has 4 lexical tones. In particular, all four Middle Chinese tones have been preserved in Standard Cantonese, in an upper and lower register, plus the Upper Entering tone (陰入) has been subdivided into two tones, for a total of 9 tones. An ongoing development further splits out an extra-high variant of the Upper Level tone (陰平), adding a 10th tone. We ignore this last distinction and further adopt a common analysis whereby the 3 Entering tones are treated as variants of other tones restricted to checked syllables (those ending in a stop consonant). This leaves us with 6 lexical tones, which cannot easily be reduced further.

These 6 tones combine nearly freely with 9 syllable nuclei

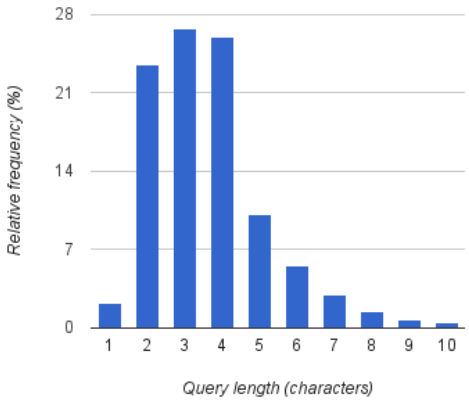

Figure 1: Length of search queries from Hong Kong.

(8 vowels plus one syllabic nasal), for a theoretical total of 54 nucleus/tone combinations. In order to limit the complexity of the resulting phonetic unit inventory, some rarely-used combinations are merged into single units, leaving us with 50 nucleus/tone units. To these we add 24 consonants and off-glides plus a silence phone for a final total of 75 phonetic units. This large inventory leads to sparseness problems in acoustic modeling.

At the level of syllables, however, this picture improves a bit. Disregarding tones, there are fewer than 700 base syllables. The total number of distinct syllables including tones is only around 2,200, meaning the effective number of tones is closer to 3 than to 6.

3.2. Lexicon

Like Mandarin, Cantonese is written in Chinese script without explicit word boundaries and has no universally agreed-upon notion of what counts as a word. Some researchers use word lists and segment input text by maximum match or more sophisticated approaches. Another option is unsupervised word segmentation where the word inventory itself is learned from data. A third and simpler approach is to use the characters themselves as basic units. Recent research [7, 8] shows that using characters-based lexicons results in good performance in Mandarin and greatly reduces system complexity.

In our Cantonese system, we also use a character-based lexicon, consisting of 9,249 Chinese characters. Each character is allowed to have multiple alternative pronunciations. Each pronunciation consists of a single syllable. In our lexicon, the average number of pronunciations per character is 1.48. The largest number occurs with the character 撱, which has 18 distinct pronunciations, most of them presumably erroneous. For comparison, our Mandarin character dictionary consists of nearly 22,000 characters and has a slightly lower average of 1.38 pronunciations per character. Just looking at the set of characters it shares with the Cantonese lexicon, the average number of Mandarin pronunciations is 1.46.

3.3. Language Model

Using characters as our basic lexical units has several advantages. First, it avoids the complication of using an arbitrary word segmentation and keeping it consistent between the language model and the lexicon. Second, it limits the size of the lexicon to a few thousand characters, versus millions of entries typically used in word-based lexicons. Finally it simplifies debugging and system design, since Chinese characters naturally correspond to an even smaller number of syllables. A potential

n-gram	# of n-gram	
	10M	70M
3-gram	69.3	56.6
4-gram	66.1	52.1
5-gram	66.0	51.3
6-gram	66.1	51.2

Table 2: Perplexities of character language models.

advantage of word-based models is that they can use more context for language modeling and need fewer pronunciations per word, since the word itself is often sufficient for disambiguating the pronunciation of its constituent characters.

To compensate for a loss of context by using character n-gram models we need to expand the order of n-grams in our language modeling. In a word-based system, trigrams might be enough to cover most contexts. In our character-based models, we are forced to expand the context to 4-grams or even 5-grams. To decide the best n-gram order for language modeling, we examined the length of search queries in Hong Kong. We only count Chinese queries and exclude English, URL, and number queries. The histogram of query lengths is shown in Figure 1. Most Chinese queries consist of 2, 3 or 4 characters, which accounts for more than 75% of all queries. The average number of characters per query is 3.65. This suggests that a 3-gram or 4-gram model may be sufficient.

We first look at the perplexities of language models with different n-gram orders. The character-based language models were trained on anonymized query logs with Chinese characters, English words, URLs, and numbers. Their perplexities are shown in Table 2. Consistent with the observations from the character histogram in Figure 1, the perplexity saturates with 4-gram order and using higher than 5-gram order yields no reduction in perplexity.

We also investigated the changes in perplexity with respect to different numbers of n-grams in Table 2. Each language model was filtered to the specified target number of n-grams using Stolcke pruning [9]. 10 million and 70 million n-grams were used as target numbers for perplexity calculation. As expected, the language model with 70 million n-grams has lower perplexity than that with 10 million n-grams in all n-gram orders. However, using more n-grams generates larger grammar FSTs and slows down decoding. Considering speed, memory usage, and performance, we decided to use a 4-gram language model with 10 million n-grams.

4. Multilingual Issues

In multilingual Hong Kong, users mix more English into their queries than users in Mainland China and Taiwan. Table 1 shows that about one third of queries in Hong Kong are in English. People in Hong Kong systematically use English words for certain consumer product, professional activities, etc.

One basic requirement is that our language models should include Chinese characters, English words, and URLs. During preprocessing, we remove all queries with non-English and non-Chinese characters. Then we add spaces around Chinese characters and leave English queries unchanged. The language models then use characters as basic units for Chinese queries and words as basic units for English queries and URLs.

Next we need to define a compact phonetic inventory that covers both Cantonese and English sounds. Since our Cantonese inventory is quite large to begin with, a way of mapping the less frequent English phonemes to existing Cantonese units is appropriate. A smaller set of units will result in better data utilization for similar pronunciations between Cantonese and English. Mapping the second language to the dominant language has been explored in other multilingual systems [10]. In the rest of this section we explore two different approaches, partial sharing and full sharing for phoneme mapping.

English lexicon:	google	g uw g ax l
Partial sharing:	google	k uw k ax l
Full sharing:	google	k u_M k 6_M l

Table 3: Partial and full sharing of non-native phonemes.

	Partial Sharing	Full Sharing
WER (%)	36.8	34.7

Table 4: Word error rates of partial sharing and full sharing for phoneme mapping on the HK testing set.

4.1. Partial Sharing

Because tones are associated with vowels, intuitively we can share consonants between English and Cantonese without any tonal issues. In this approach, we map each English consonant to a similar Cantonese consonant and leave both English vowels and Cantonese vowels in the phonetic inventory. The total number of phonetic units grows to 102. All 102 phonetic units were trained on the audio data collected from Cantonese speakers. Table 3 shows an example of partial sharing in the lexicon. In the original US lexicon "google" is pronounced as /g uw g ax l/ (ArpaBet). In partial sharing we keep vowels unchanged and map consonants to similar consonants in Cantonese as /k uw k ax l/. In this example the English phoneme /g/ is mapped to the Cantonese /k/ and English /l/ is mapped to Cantonese /l/.

4.2. Full Sharing

In order to reduce the number of phonetic units further we investigate sharing both consonants and vowels between English and Cantonese. Sharing consonants is done in the same way as in partial sharing. In addition we map English vowels to similar Cantonese vowels with the Upper Departing tone (陰去), which is mid level and hence arguably the most "neutral" of the 6 tones. English diphthongs are mapped to two consecutive vowels in Cantonese. The total number of phonetic units remains at 75, which is significant smaller than in partial sharing. The same example of "google" for full sharing is shown in Table 3 ("_M" denotes the mid level tone).

5. Experiments

5.1. Quality Metrics

In speech recognition, one typically measures word error rate (WER), which in practice often amounts to syllable error rate. Since our system contains both English words and Chinese characters, we decided to treat Chinese characters as whole words while leaving English words and other tokens (such as numbers or URLs) unsegmented. So when we report WER we are actually reporting a hybrid metric combining character error rate (CER) for Cantonese and WER for the other tokens.

Although word error rate is an important metric for speech recognition, we found it is not always helpful for developing voice search system. The overall goal of a voice search system is to help the user find information. Some users formulate queries with English words while others use Cantonese

transliterations (for example “Jordan” vs. “佐敦”). Or consider “www.google.com” and “google”: both queries yield satisfactory search results for users. However, the queries in both examples are far apart in terms of WER, obscuring the fact that they yield similar or identical results.

This makes it challenging to develop and evaluate the system, since it is often impossible for the recognizer to distinguish between an English word and its Cantonese transliteration. During development we use a metric that simply checks whether the correct search results are returned in a list of N results. This metric is called web score at N (WSC). In this paper we report WSC for $N = 1$. The top web search results produced by the recognition hypothesis are compared with the top web result produced by the reference transcript.

5.2. Data Sets

Two sets of 250k utterances each were collected in Hong Kong (HK) and Guangzhou (GZ). In each set, we randomly sampled 15k utterances for evaluation and used the remaining utterances for training. These two training sets and testing sets are further combined to form a joint training set and joint testing set. Both English and Cantonese utterances are included in all data sets. In the rest of the paper, we refer to the joint set as HK+GZ and individual sets as HK and GZ.

5.3. Phoneme Sharing

Our first experiment compares partial and full sharing for mapping between Cantonese phonemes and English phonemes. The joint HK+GZ training set of about 500k utterances was used for acoustic model training. A 4-gram language model with 10 million n-grams after Stolcke pruning was used for decoding. The HK testing set was used for evaluation and the decoder parameters were tuned for optimizing performance. The experimental results are shown in Table 4.

Using full sharing reduces WER by 1.9 percentage points compared with partial sharing. We believe this is because full sharing reduces the number of phonetic units and better utilizes the training data. If English utterances are excluded from the HK testing set, the WER is 24.4%, which is significantly better than the figures in Table 4. If only English utterances are considered, there are 5160 utterances and WER is 53.0%. Even with phoneme mapping, the system still has worse performance on English utterances than on Cantonese utterances.

5.4. Comparing HK and GZ data

The second experiment compares the HK, GZ, and HK+GZ data sets. The corresponding training sets were used for acoustic model training. A 4-gram language model with 10 million n-grams after Stolcke pruning was used for decoding. All three acoustic models were evaluated on all three testing sets and the decoder parameters were tuned for real time decoding speed. Full sharing was used for phoneme mapping. Word error rates are shown in Table 5 and web scores in Table 6.

The HK acoustic models have better performance on the HK testing set than on the GZ testing set in both word error rates and web scores. The GZ acoustic models have better performance on the GZ testing set than the HK testing set. This is because there are acoustic mistmatches between training data and testing data. Importantly, the HK+GZ acoustic models improve performance further on all three testings sets. Even with mistmatched data, more data still helps performance.

With the HK+GZ acoustic models, the performance on the HK testing set is better than on the GZ testing set. Because the language model was trained on search query logs from Hong Kong, this mismatch degrades recognition on the GZ testing set. On the HK+GZ testing set, we achieve 37.5% WER with 10.4% deletion, 6.9% insertion, and 20.2% substitution errors.

	WER (%)		
Training set	HK	GZ	HK+GZ
HK (250k)	38.3	43.6	41.3
GZ (250k)	43.8	40.1	41.7
HK+GZ (500k)	36.2	38.4	37.5

Table 5: Word error rates on different training and testing sets.

	WSC (%)		
Training set	HK	GZ	HK+GZ
HK(250k)	52.6	47.3	49.8
GZ (250k)	47.0	50.7	59.0
HK+GZ (500k)	54.7	52.9	53.8

Table 6: Web scores on different training and testing sets.

6. Conclusions

We have systematically investigated the problems of acoustic modeling, language modeling, and lexicon building for a Cantonese voice search system. With acoustic models trained on data collected from local Cantonese speakers and language models trained on search query logs we can achieve 37.5% WER and 53.8% WSC with both Cantonese and English utterances. The system was launched in December 2010 to help Cantonese speakers input text and look up information more easily. We expect the system to improve as data from the production deployment are collected and transcribed. In the future we plan to investigate better tone modeling and new approaches for dealing with English words in hybrid multilingual systems.

7. References

[1] Y. W. Wong, K. F. Chow, W. Lau, W. K. Lo, Tan Lee, and P. C. Ching, “Acoustic modeling and language modeling for cantonese LVCSR”, In Proc. of the 6th European Conference on Speech Communication and Technology, 1999.

[2] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, “Spoken language resources for Cantonese speech processing”, Speech Communication, volume 36, issue 3, 327–342, 2002.

[3] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, and M. LeBeau, “Building transcribed speech corpora quickly and cheaply for many languages”, Interspeech, 2010.

[4] M. J. F. Gales, “Semi-tied covariance matrices for hidden Markov models”, IEEE Trans on Speech and Audio Processing, 1999.

[5] “OpenFst Library, http://www.openfst.org”.

[6] D. Povey, D. Kanevsky, B. Kingsbury, and B. Ramabhadran, G. Saon, K. Visweswariah, “Boosted MMI for model and feature space discriminative training”, ICASSP, 2008.

[7] J. L. Hieronymus, X. Liu, M. J. F. Gales, and P. C. Woodland, “Exploiting Chinese character models to improve speech recognition performance”, Interspeech, 2009.

[8] J. Luo, L. Lamel, and J.-L. Gauvain, “Modeling characters versus words for mandarin speech recognition”, ICASSP, 2009.

[9] A. Stolcke, “Entropy-based pruning of backoff language models”, In Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[10] H.-A. Chang, Y.-H. Sung, B. Strope, and F. Beaufays, “Recognizing English queries in Mandarin voice search” ICASSP, 2011.