Lexical Access for Large-Vocabulary Speech Recognition

Roger H.Y. Leung and Hong C. Leung
Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
hyleung@ee.cuhk.edu.hk , hcleung@ee.cuhk.edu.hk

ABSTRACT

In this paper, the lexical characteristics of two Chinese dialects and American English are explored. Different lexical representations are investigated, including the tonal syllables, base syllables, phonemes, and the broad phonetic classes. Multiple measurements are made, such as coverage, uniqueness, and cohort sizes. Our results are based on lexicons of 44K and 52K words in Chinese and English obtained from the CallHome Corpus and the COMLEX Corpus, respectively. We have found that the set of the most frequent 4,000 words has coverage of 92% and 77% for Chinese and English, respectively. The phonetic representation unique specifies 85%, 87% and 93% of the lexicon for Mandarin, Cantonese, and English, respectively. While the three languages appear quite different when they are described by their full phoneme sets, their characteristics are more similar when they are represented in terms of broad phonetic classes.

1. INTRODUCTION

Chinese is different from many western languages in that it is monosyllabic and tonal. While there are more than 15,000 monosyllabic Chinese characters, there are typically only about 1,300 tonal syllables in each of the Chinese dialects. Thus, many Chinese characters share the same pronunciation. However, often depending on the context, each Chinese character may have multiple pronunciations. As a result, Chinese is a complex language with many-to-one and one-to-many mappings between the characters and the syllabic pronunciations. The notion of a Chinese word is also very different from many western languages. While the syllables and characters are relatively well defined, the Chinese words are composed of a variable number of characters. Since a Chinese word can be formed, in principle, by any combination of ~15,000 Chinese characters, the vocabulary of a speech recognition system can be huge. Hence the computation time for speech recognition can become extremely expensive.

In this paper, we will investigate the lexical characteristics of two of the most commonly spoken Chinese dialects: Mandarin and Cantonese, and compare the results with American English. The relative uniqueness, coverage, and cohort sizes of the languages will be examined. Partial phonetic description will also be explored.

Lexical access from partial phonetic information has been proposed for American English [1]. Rather than performing detailed phonetic analysis, a word is characterized in terms of broad phonetic classes. This partial description is then used to retrieve a small set of words from a large lexicon. Our lexical study is based on Mandarin CallHome and English COMLEX obtained from the Linguistic Data Consortium [2,2]. The Chinese lexicon is consisted of 44,000 words, and the English lexicon is consisted of 52,000 words.

The six broad phonetic classes are formed based on manner of articulation. They are "*vowels, stops, fricatives, affricates, laterals/glide, nasals" for* Chinese dialects and *"vowels, stops, strong fricatives, weak fricatives, laterals/glide, nasals"* for English. This set of manner classes is used, since they tend to be relatively invariant across different speakers and phonetic contexts.

2. METHODOLOGY

Each of the lexicons for the three different languages is represented in multiple units, including tonal syllables, base syllables, phonemes, and broad phonetic classes. In order to explore the characteristics of the languages, multiple measurements are made, such as coverage, uniqueness, expected and maximum cohort sizes. Since words in a lexicon may have very different frequency of occurrence, some of our measurements are also weighted by the frequency of occurrence. The frequency of occurrence for English is obtained from the Brown Corpus, whereas the frequency of occurrence for Mandarin and Cantonese are obtained from the CallHome database.

Table 1 shows some of the basic measurements used in our study. The maximum cohort size represents the largest equivalence class size given a particular phonetic / syllabic description, whereas the expected cohort size represents the cohort size with a frequency distribution. Notations for different measurements are shown in Table 2.

	Uniform Distribution	Frequency Normalized
Maximum cohort size	$\max_{w_i \in L} \lvert C(w_i) \rvert$	$\max_{w_i \in L} \lvert C(w_i) \rvert$
Expected cohort size	$\frac{1}{\lvert L \rvert} \sum_{w_i \in L} \lvert C(w_i) \rvert$	$\sum_{w_i \in L} p_i \lvert C(w_i) \rvert$

Table 1: The basic measurements used in our study. $|C(w_i)|$ is the cohort size for word w_i, $|L|$ is the lexicon size, and p_i is the frequency of occurrence of the i'th word, w_i,, in lexicon L.

Notation	Statistics
UNIQ	% of word which is uniquely specified
ECS	Expected cohort size
F-ECS	*Frequency normalized expected cohort size*
MCS	Maximum cohort size
RECS	Expected cohort size /lexicon size
F-RECS	*Frequency normalized expected cohort size /lexicon size*
RMCS	Maximum cohort size /lexicon size
LEX	Lexicon size

Table 2: Notations for the measurements used in this study. Results normalized by frequency of occurrence are shown in *italic.*

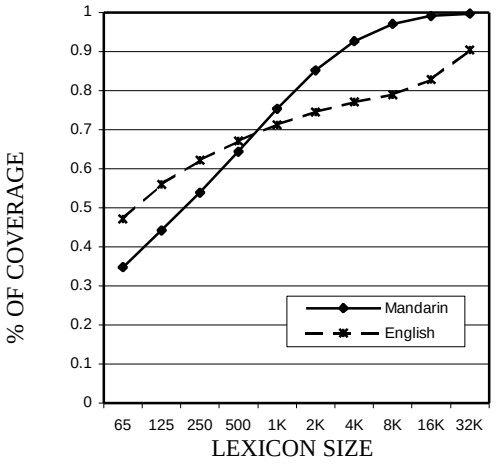

Figure 1: Percentage of text coverage for English and Mandarin most frequent words

3. EXPERIMENTS

In all languages, some words occur much more frequently than others. Therefore, it would be interesting to see the frequency distribution of the words in a language. Figure 1 shows the cumulative distribution of the most frequent words for Mandarin and English. For example, the set of the most frequent 4,000 words cover over 92% and 77% of all the texts in CallHome Mandarin and the Brown Corpus, respectively.

The characteristics of Mandarin are studied by representing the Mandarin lexicon in terms of tonal syllables, base syllables, and 6 broad classes. The experimental design is similar to that of Huttenlocher [1]. Table 3 shows our results for Mandarin. It can be seen that if the lexicon is represented in terms of the tonal syllables, only 85% of the lexicon can be uniquely specified. The remaining 15% of the lexicon contain words that cannot be uniquely specified by the tonal syllables. This low percentage of 85% reflects the fact that many of the words in Mandarin are actually homophones. For example, all of the following Chinese words have the same tonal-syllable representation, "*fu4 shu4*":

復述	複數	負數	富庶

When the lexicon is represented in terms of the base syllables, i.e. syllables with no tone information, only 65% of the lexicon can be uniquely specified. Similarly, only 19% of the lexicon can be uniquely specified by the broad classes.

However, Table 3 also shows the discriminatory power of the broad phonetic classes. By using only 6 broad phonetic classes, the expected cohort size (ECS) is found to be 62.4. In other words, if the lexicon is represented in terms of the broad classes, on average 62.4 words would have the same broad class representation.

	Tonal Syllable	Base Syllable (38 phoneme)	Manner of Articulation
UNIQ	85.0%	65.0%	19.0%
ECS	1.39	2.54	62.4
F-ECS	*3.44*	*9.24*	*127.6*
MCS	21	54	299
LEX	44K	44K	44K

Table 3: Analysis on Mandarin broad classes

Table 4 shows the characteristics of Cantonese. We can see that 87%, 70%, and 16.7% of the lexicon can be uniquely specified by the tonal syllables, base syllables, and broad phonetic classes, respectively. These figures are quite similar to those for Mandarin.

However, the expected cohort size in Cantonese is 107.9, almost twice of the corresponding size in Mandarin. This shows that the broad phonological structures for the two Chinese dialects are quite different. It also suggests that the 6 broad phonetic classes are not as effective in differentiating the Cantonese words in the lexicon.

	Tonal Syllable	Base Syllable (38 phoneme)	Manner of Articulation
UNIQ	87.2%	70.1%	16.7%
ECS	1.32	2.15	107.9
F-ECS	*2.69*	*6.80*	165.9
MCS	26	37	471
LEX	44K	44K	44K

Table 4: Analysis on Cantonese broad classes

Table 5 shows our analysis for English. It can be seen that over 93% of the lexicon can be uniquely specified by a set of 43 phonemes, in contrast to the 85% and 87% for Mandarin and Cantonese with tonal information. Furthermore the expected cohort size is about 74, which is comparable to the corresponding figures in Mandarin and Cantonese. These experimental results for English are very similar to those reported by Carter [5]. We have found that the largest broad class cohort is: [fricative] [Vowel] [fricative] [vowel] [fricative]. This cohort has 648 word members, such as "thesis", "visit".

	43 Phonemes	Manner of Articulation
UNIQ	93.2%	15.7%
ECS	1.07	74.1
F-ECS	*1.83*	*111.5*
MCS	5	648
LEX	52K	52K

Table 5: Analysis on English broad classes

In order to compare directly the lexical characteristics of all three languages, Table 6 summarizes the results when the base syllables (or 38 phonemes) are used for the Chinese dialects, and the set of 43 phonemes is used for English. It can be seen that the characteristics of the three languages are quite different. First, there is a major difference between the UNIQ's for the three languages, ranging from 65% for Mandarin to 93% for English. Second, the relative cohort sizes can differ by as much as a factor of 2.7, since the RECS for Mandarin is 0.0057% and the RECS for English is 0.0021%. Finally, the RMCS can also differ by an order of magnitude, since the RMCS for Mandarin is 0.12% and the RMCS for English is 0.0096%.

We have also compared the lexical characteristics of the three languages using the 6 broad classes. Table 7 summarizes the results. We can see that their characteristics are more similar than those using the entire phoneme set. First, it is observed that almost 20% of the Mandarin lexicon can be uniquely defined by the broad phonetic classes, compared to 15.7% for English. Second, the relative expected cohort sizes are quite small for all three languages, with the highest one at 0.24% for Cantonese and the lowest one at 0.14% for both Mandarin and English. Third, while the maximum class sizes for all three languages are still quite low, they differ by only a factor of 2. For example, the RMCS for Mandarin is 0.67%, whereas that for English is 1.25%.

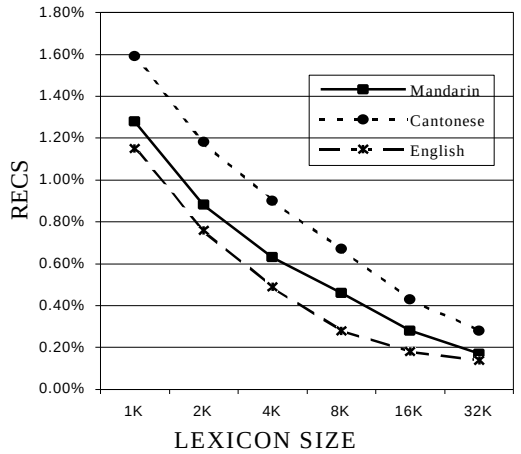

Figure 2: Relative expected cohort size analysis of 6 broad classes.

The effectiveness of the broad class representation for the three languages are compared, Figure 2 shows the relative expected cohort sizes (RECS) as functions of the lexicon sizes. It can be seen that the RECS decrease monotonically. With a lexicon size of 4,000, the RECS for all languages are below 1%.

Figure 3 to Figure 6 compare the characteristics of Mandarin and English as functions of the lexicon sizes. We can see that most of the curves are quite linear with the lexicon size and that the characteristics using broad phonetic classes are quite similar between the languages.

	Mandarin	Cantonese	English
UNIQ	65%	70.1%	93.2%
ECS	2.54	2.15	1.07
MCS	54	37	5
RECS	0.0057%	0.0049%	0.0021%
RMCS	0.12%	0.083%	0.0096%
LEX	44K	44K	52K

Table 6: Comparisons of characteristics between Mandarin, Cantonese, and English. Both Mandarin and Cantonese are based on the base syllables (or 38 phonemes), whereas English is based on a set of 43 phonemes.

	Mandarin	Cantonese	English
UNIQ	19.0%	16.7%	15.7%
ECS	62.4	107.9	74.1
MCS	299	471	648
RECS	0.14%	0.24%	0.14%
RMCS	0.67%	1.1%	1.25%
LEX	44K	44K	52K

Table 7: Analyses on Mandarin, Cantonese, and English for six broad classes

5. SUMMARY

In summary, we have described and compared the lexical characteristics of three languages: Mandarin Chinese, Cantonese Chinese, and English. While the three languages appear quite different when they are described by their full phoneme sets, their characteristics are more similar when they are represented in terms of broad phonetic classes.

6. REFERENCES

1. D.P. Huttenlocher & V.W. Zue, "Phonotactic and Lexical Constraints in Speech Recognition", Proc. American Association for Artificial Intelligence Conference, pp.172-176, 1983.

2. Huang, Shudong, X. Bian, G. Wu, and C. McLemore, "LDC Mandarin Lexicon", LDC, Univ. of Pennsylvania, Philadelphia: Linguistic Data, 1997.

3. R. H. Baayen, R. Piepenbrock & L. Gulikers, "The CELEX Lexical Database (CD-ROM)", LDC, Univ. of Pennsylvania, Philadelphia, PA, 1995.

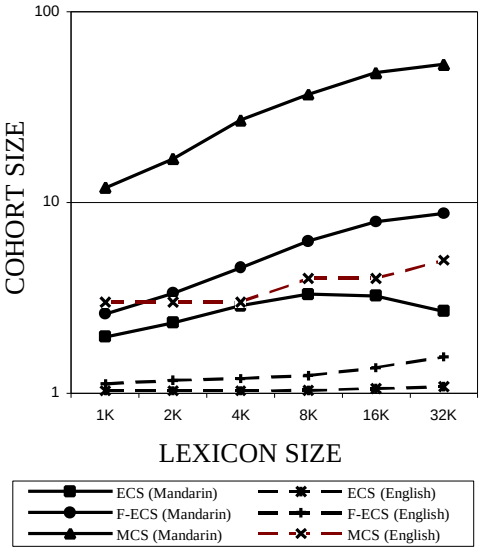

Figure 3: Expected and Maximum Cohort Size for whole phoneme set

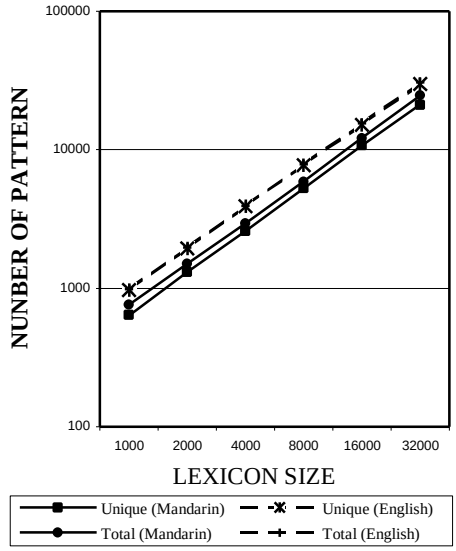

Figure 4: Total number of pattern and number of unique pattern for whole phoneme set

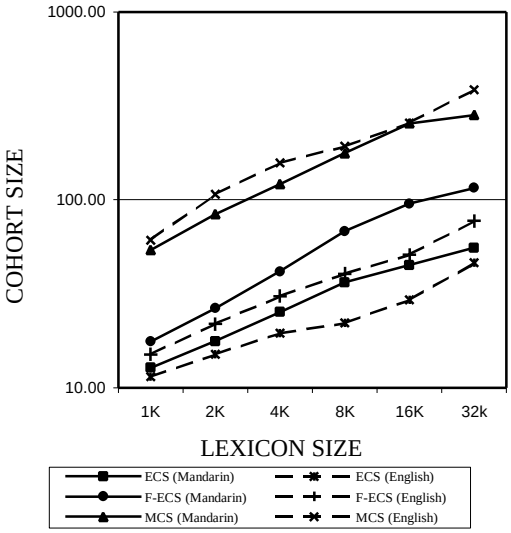

Figure 5: Expected and Maximum Cohort Size for 6 broad classes

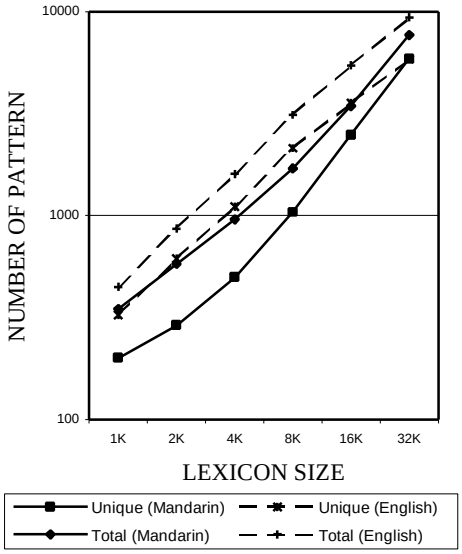

Figure 6: Total number of pattern and number of unique pattern for 6 broad classes

4. J.X. Wu, Li Deng, Jacky Chan, "Modeling Context-Dependent Phonetic units in a Continuous Speech Recognition System for Mandarin Chinese", International Conference on Spoken Language Processing,pp.2281-2284, 1996.

5. David M. Carter, "An information-theoretic analysis of phonetic dictionary access", Computer Speech and Language, pp.2,2-11, 1987.

6. D.W. Shipman, & V.W. Zue, "Properties of Large Lexicons: Implications for Advance Isolated Word Recognition System", Proc. IEEE ICASSP pp.546-549, 1982.

7. Jialu Zhang, "Phonetic and Linguistic Features of Spoken Chinese", International Symposium on Speech, Image Processing and Neural Networks pp.117-121, 1994.

8. P. Ladefoged, "A course in Phonetics, 3 rd edition", Harcourt Brace Jovanovich, 1993