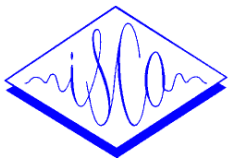
ISCA Archive
http://www.isca-speech.org/archive



3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary

Yen-Ju Yang[1], Sung-Chien Lin[1], Lee-Feng Chien[2], Keh-Jiann Chen[2], and Lin-Shan Lee[1,2,3]
[1]Dept. of Computer Science and Information Engineering, National Taiwan University
[2]Institute of Information Science, Academia Sinica
[3]Dept. of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

Abstract

This paper proposes a word-class-based Chinese language model for Mandarin speech recognition with very large vocabulary. The word classes used are developed based on the special structure of Chinese words. We have also developed some improved techniques. The ambiguous syllable filter can delete many confusion syllables and increase significantly the accuracy. The short-term cache memory can help the language model to adapt to the current application domain, and the learning module can significantly reduce the zero values in the language model.

1. Introduction

Today, the input of Chinese characters into computers is still a very difficult and unsolved problem. This is the basic motivation for the development of Mandarin speech recognition techniques with very large vocabulary. Because of the monosyllabic structure of Chinese language, i.e., every Chinese character is pronounced as a monosyllable, it is believed that monosyllabic-based approach is currently the most feasible for this recognition task. In this approach, the input speech is a sequence of monosyllables, so each monosyllable is first recognized based on acoustic features as a set of top n candidates due to the high degree of confusion in these syllables, and a Chinese language model is then used to find the most promising character each syllable actually represents, because very often many homonym characters share the same syllable pronunciation.

The basic structure of such a system is shown in Fig. 1 [1].

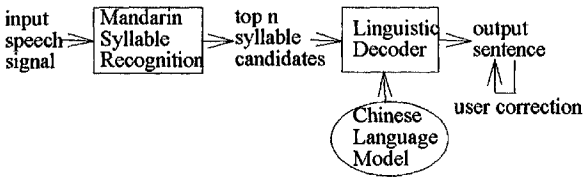


Fig. 1 The overall structure of a monosyllable-based Mandarin speech recognition system with large vocabulary

This paper presents an intelligent and efficient Chinese language model used in such an approach. In this approach, all possible word hypotheses are first obtained to construct a Chinese word lattice, and then a word-class-based bigram Markov model with several improved techniques integrated is used to select the most promising concatenation of word hypotheses as the output sentence. The improved techniques includs ambiguous syllable filtering, short-term cache memory, long-term learning and fast table matching as shown in Fig. 2. With these techniques high correct rate, high adaptation flexibility and high speed of processing can be achieved.

In the experiments, the lexicon contains 84,495 words, and the training corpora contain three months of newspapers, some articles from magazines and parts from various novels, with a total of 5,303,554 characters (3,500,067 words). The test texts contain pieces of news, magazine article and tales, with a total of 3,733 characters (2,394 words).

The remainder of this paper is organized as follows : Section 2 describes the Chinese word-class-based bigram Markov model and the baseline experiments, Section 3 presents the improved techniques and experimental results,Section 4 is the concluding remarks.

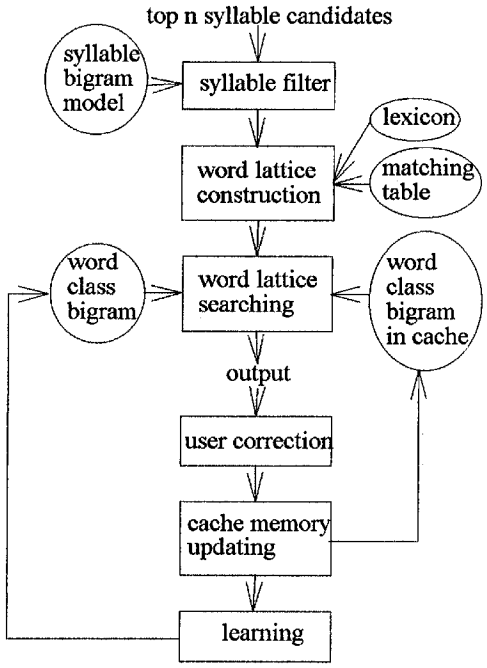


Fig.2 The word-class-based Chinese language model with improved techniques proposed in this paper

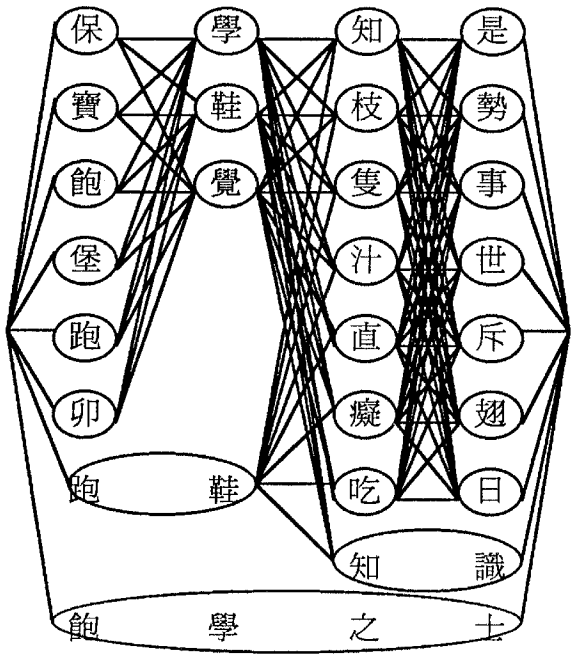


Fig. 3 A simplified word lattice

2. Word-Class-Based Chinese Language Model

In Chinese language, every character is monosyllabic but every word is composed of from one to several charaters. Thus, the recognized syllable strings are first matched with the words in a lexicon to find all possible word hypotheses to construct a word lattice. A word lattice is a graph of all possible pathes connecting all word hypotheses, a good example is in Fig. 3. An output Chinese sentence can then be found by concatenating the words in the lattice.

In language modeling a large vocabulary implies high degree of ambiguity,high computation load and large memory requirements. More importantly, it will be difficult to estimate, store and retrieve large number of statistical parameters for a language model with a large vocabulary. These problems are very serious in the case here, because there exist about 85,000 words in the lexicon, which requires 85,000x85,000 parameters for a word bigram. However, considering the special structure of Chinese language, a nice observation is that many words with similar linguistic properties have the same ending characters, e.g. "我們 (/wo/ /men/, we)", "你們 (/ni/ /men/, you)" and "他們 (/ta/ /men/, they)", while many others with similar linguistic properties have the same starting characters, e.g. "喜歡 (/shi/ /huan/, like)", "喜愛(/shi/ /ei/, love)" and "喜好 (/shi/ /hao/, prefer)" . It is therefore computationally straightforward to group words into class according to the ending/starting characters, i.e. all the words with the same ending/starting character form a class. Based on such word classes a word-class-based Chinese language model is then proposed here, which can reduce the required parameters from 85,000x85,000 to about 13,000x13,000 [2].

After a word lattice was constructed as discussed above, the proposed word-class-based Chinese language model was used to search through the word lattice to obtain the maximum likelihood output sentence. For each word hypotheses sequence $W=w_1w_2...w_m$, where w_i is the i-th component word hypothesis, let $S(w_i)$ and $E(w_i)$ be the word class with starting and ending characters identical to those of w_i, then

$$P(w) = P(w_1, w_2, \ldots, w_m)$$
$$\approx P(w_1)p(w_2|w_1)\ldots p(w_m|w_{m-1}) \quad (1)$$

$$P(w_i|w_{i-1}) \approx P(w_i|S(w_i))P(S(w_i)|E(w_{i-1}))P(E(w_{i-1})|w_{i-1})$$
$$= P(w_i|S(w_i))P(S(w_i)|E(w_{i-1}))$$
$$\because P(E(w_{i-1})|w_{i-1}) \quad (2)$$

Although the number of statistical parameters of this language model has been significantly reduced, very probably many zeros still exists constrained by the training corpus. So the smoothing technique is very important, we modified the nonlinear interpolation method [3] as follow :

$$P(w_i|S(w_i)) = \frac{\max\{N(w_i) - d, 0\}}{N(S(w_i))} + d\frac{N(N(w) \geq d)}{N(S(w_i))} * \frac{1}{L(S(w_i))} \quad (3)$$

$$P(S(w_i)|E(w_{i-1}))$$
$$= \frac{\max\{N(E(w_{i-1}), S(w_i)) - d, 0\}}{N(E(w_{i-1}))} + d\frac{N(N(E(w_{i-1}), S(w_i)) \geq d)}{N(E(w_{i-1}))} P(S(w_i))$$
$$P(S(w_i))$$
$$= \frac{\max\{N(S(w_i)) - d, 0\}}{\sum_{w \in lexicon} N(S(w))} + d\frac{N(S(w)) \geq d)}{\sum_{w \in lexicon} N(S(w))} * \frac{1}{L} \quad (4)$$

where N(<x>) is the number of <x> occurring in the training corpus, L is the number of words in the lexicon and L(S(w_i)) is the number of distinct words in the same word class S(w_i).

The performance of the word-class-based language model has been compared with that of the word bigram language model as in Table 1. The training and testing database is described in Section 1, and both models accept top 5 syllable candidates as input. From Table 1, we can find taht the character correct rate of the proposed language model is higher than the word bigram language model.

model \ test	1	2	3	4	5	Avg
WB	78.93	76.43	85.21	71.38	67.42	75.87
WCB	89.36	83.01	85.46	81.98	79.28	83.82

Table1 The character correct rate(%) of word bigram(WB) and the proposed word-class-based bigram(WCB), 1 to 5 are five differnt sets of testing texts

3. Improved Techniques

The Chinese language model works on the top n syllable candidates recognized by acoustic features, therefore the syllable recognition will significantly affect the final correct rate of the language model as shown in Fig. 4. It is cleared that the more number of syllable candidates are included, the higher the overall syllable recognition rate is, but the final correct rate for characters is not necessarily proportion to the overall syllable recognition rate, because large number of syllable candidates apparently result in high degree of ambiguity among word hypotheses.

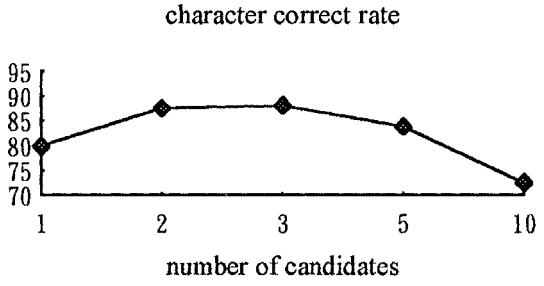


Fig. 4 Character correct rate between different number of candidates

We thus must try to keep the high overall syllable recognition rate but in the same time small number of syllable candidates. An ambiguous syllable filter is therefore constructed as a preprocessor for the language model to delete many impossible syllable hypotheses selected in the acoustic recognition process as the top n candidates. This syllable filter works as a communication channel. It reestimats the scores of each syllable and delete many impossible syllable hypotheses with low scores.

Adaptation capabilities are very important in large vocabulary speech recognition, because when dictating some texts on some topics, some special word patterns used in the immediately past are very likely to be used again soon. A short-term word-class-based language model maintained in the cache memory is therefore developed [4]. This short-term cache memory can help the system to adapt to the current text and application domain and significantly improve

the on-line performance. When we search through the word lattice to obtain the maximum likelihood output sentence, we first retrieve statistical parameters from the short-term cache memory. If it is zero in the cache memory then we retrieve it from the normal language model.

Because for our system the input sentence can be of arbitrary unlimited texts, so the statistical parameters are very often insufficient because actually there exist many word patterns which have never occurred in the training corpus. In order to solve this problem in our language model a learning module is integrated, which can continuously update all the statistical parameters according to the input sentences. In other words, the language model is not at all static, it is dynamically updating through learning. The more frequently the system is used, the more reliable and robust the language model becomes. With this learning capabilities integrated, the performance is further improve, as can be found in Fig. 5.

Real time processing is an important

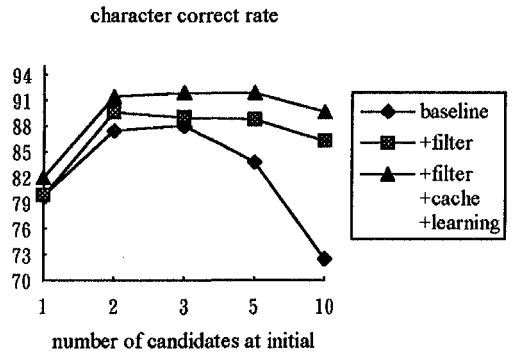


Fig. 5 The character correct rate with improved techniques

requirement for practical speech recognition system. It has been observed that the Chinese language model spend most of the CPU time on word hypotheses formation during constructing the word lattice, because of the very large number of words in the exicon. If this matching time can be reduced, it will help speed up the response time. Based on the syllable sequences in the words in the lexicon, two bit tables, the relative position table and absolute position table, are developed to speed up the word hypotheses formation processes. With these two tables, about 70% of the number of matching operations can be eliminated.

4.Concluding Remarks

The word-class-based Chinese language model discussed in this paper has been succeessfully implemented first on a SUN workstation and then further incorporated with a PC-based real time Mandarin dictation machine [1]. Very satisfactory performance has been observed, and the functions of all the improved techniques mentioned here have been tested and proved. It is cleared that the proposed model is very efficient for Mandarin speech recognition with very large vocabulary.

Acknowledgement

The authors would like to acknowledge the indispensible contribution made by the Chinese Knowledge Information Processing Group in Institute of Information Science, Academic Sinica and Prof. Ruei-Tuen Wu of Dept. of Psychology, National Taiwan Universaity for their invaluable help in providing experimental corpora.

[Reference]

[1] Lee, L. S., et al., "Golden Mandarin (II)-an Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," Proc. *ICASSP*93, pp. II503-II506, 1993

[2] Lin, S. C., Chien, L. F., Chen, K. J., Lee, L. S., "A Word-Class Bigram Approach to Linguistic Decoding in Mandarin Speech Recognition," Proc. *ROCLIN*VI, 1993

[3] Ney, H., Essen, U., "On Smoothing Techniques for Bigram-Based Natural Language Modeling," Proc. IEEE *ICASSP*91, pp. 825-828, 1991

[4] Kuhn, R., Mori, R.D., "A Cache-Based Natural Language Model for Speech Recognition," IEEE Trans. PAMI-12, no. 6, pp. 570-583, Jun. 1990