

## Chinese Speech Understanding and Spelling-Word Translation Based on the Statistics of Corpus

Jun WU, Zuoying WANG, Jiasong SUN, Jin GUO

Department of Electronic Engineering, Tsinghua University  
Beijing, 100084, China

### Abstract

In this paper, a new natural language processing approach based on the statistics of corpus is proposed and has been successfully applied to THED-919 Chinese speech recognition system to eliminate acoustic recognition errors and to translate spellings into Chinese words. The accuracy rate of spelling-word translation of unrestricted text is 98.4% and 2/3 of acoustic recognition errors are eliminated.

**Keywords:** Speech Recognition, Speech Understanding, Markov Chain, N-Gram Model.

### Introduction

In terms of speech recognition, Chinese has two major differences from alphabetic languages. Firstly, all Chinese words are monosyllabic and the number of syllables is limited (about 1300); secondly, each Chinese phonetic sound (in the alphabetic representation of Chinese syllables) maps many Chinese words (about 6 Chinese words (Kanji) in common). Therefore, Chinese speech recognition has two obvious features--the acoustic recognition is easier than that of alphabetic languages, but the speech understanding is much more difficult. As a major part of any Chinese dictative machine, the speech understanding unit must not only eliminate acoustic recognition errors, but also translate spellings into words, so that this unit is more important to Chinese speech recognition than to that in any other alphabetic language.

Although the research on Chinese understanding has been done for more than a decade, the major approaches are grammar based, which are successful only in vocabulary and domain restricted speech understanding problems, but could hardly be used in the understanding of unrestricted text. It is obvious that these approaches are

not practical for a dictative machine that would recognize all Chinese words.

This characterization suggests that a solution to this problem relies on an alternative to the grammar-based approach. The approach presented in this paper is based upon considering a word sequence in a real text as a stochastic process, and applying the stochastic language model to parse this sequence. The language model is trained by a very large corpus covering most of language phenomena and can be used to solve the problem of unrestricted natural language text understanding.

Our goal is to develop a dictative machine that can accept unrestricted Chinese text without keyboard, so practicality of the language model is very important and both accuracy and speed must be considered. Although several Chinese stochastic language models have been presented, which are word based or even phrase-based[1,2,3], they do not meet the need of real time speech recognition because they can not be realized on real time mode on an entry-level computer. Fortunately, it has been found that spelling level understanding is very effective[4,5,6], so we propose a new idea of speech understanding based on the spelling level.

We present here a two-staged language model which is a hybrid of a spelling based model and a phrase-based model. Erroneous hypotheses of spelling given by acoustic recognizer should be filtered by the speech understanding unit based on the spelling level first, so each sound uttered maps very few spelling hypotheses (no more than 4) and then Chinese spelling sequences are translated into word sequences by the phrase-based model. This approach enables the system to perform effectively in both speed and accuracy, and has been successfully applied to THED-919 Chinese speech recognition system to eliminate acoustic recognition errors and to translate spellings into Chinese words. The accuracy rate of spelling-word translation of unrestricted

text is 98.4% and 2/3 of acoustic recognition errors are eliminated.

We will describe our two-staged language model in section II, and training methods and corpus in Section III. The experimental results are introduced in section IV. The last section gives the conclusion and our further research interests.

### Two-Stage Bigram Model

The goal of speech recognition by the statistic approach is to pick out a most likely sentence (word string) according to the observed acoustic evidence. Let S denotes the result of a sentence of M words where A is its observed evidence. It is obvious that:

$$S = \text{Arg Max}_j P(S^{(j)}|A)$$

$$= \text{Arg Max}_j P(A|S^{(j)}) \cdot P(S^{(j)}) \quad (1)$$

where j is the hypotheses' order. The probability  $P(A|S^{(j)})$  is estimated by the acoustic model and  $P(S^{(j)})$  could be given by language models which could study the context of a sentence. We regard a sentence as a N-1 scaled Markov chain and introduce N-1 null words at the beginning of it, then the probability  $P(S^{(j)})$  may be approximated by:

$$P(S^{(j)}) = P(W_1^{(j)}, W_2^{(j)}, \dots, W_M^{(j)})$$

$$= \prod_{i=1}^M P(W_i^{(j)}|W_1^{(j)}, \dots, W_{i-1}^{(j)}) \quad (2)$$

where  $S^{(j)} = (W_1^{(j)}, W_2^{(j)}, \dots, W_M^{(j)})$  and  $W_i^{(j)}$  is the basic unit of understanding and i is the time order.

The model above is a N-Gram model, and the model we presented in this paper is bigram model where N=2. This kind of stochastic models can be used to increase acoustic recognition rate and translate spelling sequences into word sequences simultaneously. In Chinese speech understanding, the basic unit of the bigram model could be spelling, word or even phrase. All the three kinds of models are very efficient in eliminating acoustic recognition errors, but the model based on spelling needs the least computation and memory space, and it works hundreds of times faster than the phrase-based model. Unfortunately it can not accomplish the translation from spellings to words, so when spelling-word translation is concerned, the word-based or phrase-based model must be used, and the phrase-based model is better than the word-based one in accuracy, although it requires more computation and space. Hence, there are at least three kinds of models which can be used to Chinese speech understanding, including:

1. the word based model that eliminates errors and translates spellings into words simultaneously,
2. the phrase based model that eliminates errors and translates spellings into words simultaneously,
3. the two-staged model which is a hybrid of a spelling-based model and a phrase-based model, while the former one filters most of the unlikely hypotheses of spelling from all acoustic recognition results and then the latter one gives out the sentence from the remaining hypotheses.

The first approach needs the least resource of computer but its accuracy is not very satisfying. The latter two approaches are almost the same in accuracy[4,6], but the third one is much more faster than the second one because it takes the advantage of both the spelling-based model and the phrase-based model. We have applied this model to our speech recognition and understanding system, the data flow chart is:

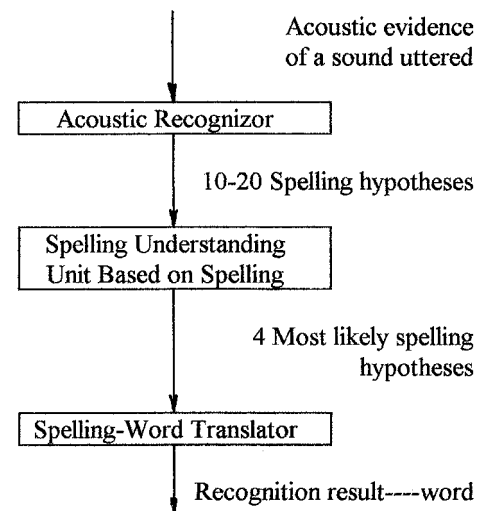


Figure 1. Data flow chart of THED-919

The key to build a language model is to estimate the probabilities  $P(W_i|W_{i-1})$ . They can only be obtained through the statistics of large corpus, and the performance of the model depends entirely on training. Two critical problems of training are the selection of training data and the estimation of probabilities from sparse data.

### Corpus and Training

#### Vocabulary

The system contains all Chinese GB (Chinese National Standard ) words (Kanji), whose number is 6724.

The phrases in the system vocabulary are selected from [8,9], and the longest phrases contains 6 words. The distribution of phrase length is:

Phrase len.	1	2	3	4	5	6
Num.	6724	30151	5125	13392	382	145

Table 1. Distribution of phrase length

The system also allows the user to build an additional vocabulary as a complement to the system vocabulary.

### Corpus

Our corpus is made up of one-year news dispatches from Xinhua News Agency, which represent about 20,000,000 words. The reason that we select this kind of corpus is that the news dispatches cover all kinds of domains, and have all kinds of sentence patterns and writing styles. We used a half of it as training data, and the other half as test data.

### Probability Estimation

The probability  $P(W_i|W_{i-1})$  could be estimated by the simple relative frequency approach according to the statistics of large corpora, and we used deleted interpolation to smooth zero probabilities [7,6], i.e.:

$$P(W_i|W_{i-1}) = \lambda_2 \cdot f(W_i|W_{i-1}) + \lambda_1 \cdot P(W_i) \quad (3)$$

$$f(W_i|W_{i-1}) = \frac{N(W_{i-1}, W_i)}{N(W_{i-1})} \quad (4)$$

$$\lambda_1 + \lambda_2 = 1, \quad 0 \leq \lambda_1, \lambda_2 \leq 1$$

and  $\lambda_1, \lambda_2$  can be got by experiments.

### Pre-Training---Phrase boundary parsing

The boundary of phrases in corpus must be known before the corpus is used to train a phrase based N-gram model. Because the size of the corpus is too large, boundary parsing could only be done by a automatic parser. We also use the bigram model to deal with this problem.

Supposing  $S=(C_1, C_2, \dots, C_L)$  is the sentence to be parsed and  $C_1, C_2, \dots, C_L$  are words in the sentence. The goal of getting phrase boundary is to find a phrase sequence  $(W_1, W_2, \dots, W_M)$  which,

$$W_1 = C_1, \dots, C_{m_1},$$

$$W_2 = C_{m_1+1}, \dots, C_{m_2},$$

.....

$$W_M = C_{m_{M-1}+1}, \dots, C_L,$$

with maximum probability  $P(W_1, W_2, \dots, W_M)$  where

$$P(W_1, W_2, \dots, W_M) = \prod_{i=1}^M P(W_i|W_{i-1}) \quad (5)$$

It is obvious that formula (5) can be estimated only when  $P(W_i|W_{i-1})$  is known. The boundaries of phrases in a subset of training corpus must be given artificially first, so an approximate estimation of  $P(W_i|W_{i-1})$  could be obtained by the statistics of that subset. Then a much larger subset can be parsed automatically, and the errors are determined artificially. The probability could be re-estimated with the parsed corpora. We continue this procedure several times, and the accuracy rate of boundary parsing can be as high as 99.9% [3].

### Experimental Results

The experimental results of understanding unit and the whole system (THED-919) are introduced respectively in this paper.

#### Spelling Understanding

When spelling based bigram model is introduced, 2/3 of spelling errors are eliminated [4,6].

#### Spelling-Word Translation

The test corpora, which consist of about 1,500,000 words and cover many areas including economy, politics, diplomacy, education, science, commerce, athletics and daily life etc., are selected randomly from corpora not included in the training set. The texts are translated into spelling sequences first and then spelling sequences are translated back into words by using our phrase-based bigram model. The results are listed in the following table:

Unit	Total Num.	Correct Num.	Accuracy Rate
Word	1,524,324	1,518,987	98.48%
Sentence	87,934	77,927	88.62%

Table 2. Accuracy of speech understanding

These results are much better than those of [3], due to the smoothing technology.

### Overall performance of the system

The test data is obtained by recordings of 3 male and 2 female. The acoustic recognition result of each sound uttered gives out 10 - 20 hypotheses. Then we use the two-staged bigram model to get the Chinese word result. It is obvious that the accuracy of word and acoustic recognition rate are in direct proportion.

User No.	Accuracy of 1 Spelling Hypothesis	Accuracy of 12 Spelling Hypotheses	Accuracy of Word after Understanding
F1	83.1%	97.5%	92.9%
F2	88.4%	98.3%	94.8%
M1	77.6%	96.2%	91.7%
M2	87.9%	98.5%	94.7%
M3	91.8%	99.1%	96.2%

Table 3. Word accuracy of different speaker

Table 3 shows that the THED-919 system is practical in accuracy when our speech understanding approach is applied.

### The Speed

The experiment is done on 80X86 PCs, the speeds of our approach are:

Computer	80386/33 without 387	80486/50
Recognition Speed	>120	>150
Understanding Speed	105	490
System Speed	>80	>110

Table 4. System speeds of THED-919 (words/min.)

If we only use the phrase-based bigram model instead of the two-staged bigram model, the understanding speed reduce to 1/8 of that of the latter.

### Conclusion and Future Work

The experimental results show that the approach based on the statistics of corpus is very effective in a unrestricted speech recognition system, and the two-stage bigram model is very practical in both speed and accuracy.

Because the training data are always insufficient, the smoothing method is very important for unrestricted texts understanding.

We will apply a trigram model based on spelling to our two staged model, since it is also very fast. We have found that the trigram model is more effective[6].

### Acknowledgment

The authors are grateful for many suggestions they have received from professor Dajin LU, professor Changning HUANG and professor Xing LI at Tsinghua University. They also wish to express their thanks to Dr. Tianying JI, Mr. Xi Xiao and Ms. Xia Wang for many help they offered.

### Reference

- [1] H.Y. Gu, et al., Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters, *Computer Speech and Language*, Vol.5,N.4, pp.363-377, 1991
- [2] L.S. Lee, et al., A Mandarin Chinese Machine Based Upon a Hierarchical Recognition Approach and Chinese National Language Analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.12, N.7, pp. 695-704, Jul. 1990
- [3] J. Guo et al., A New Approach of Analyzing Modern Chinese based on Corpus and a New Spelling-Word translator, *The 1st National Conference of Computer Linguistics*, Nov. 1991. Hangzhou China.
- [4] J. Wu, Research and Implementation of Chinese Speech Understanding Methods Based on Chinese Spelling, M.S. Degree Thesis Tsinghua University, 1993
- [5] J. Wu, Z. Wang, A New Approach of Speech Understanding Based on Corpus, *The 2st National Conference of Computer Linguistics*, pp96-101, Nov. 1993. Xiamen China.
- [6] J. Wu, Z. Wang, Y. Ren, Stochastic Language Models for Chinese Speech Recognition Based on Chinese Spelling., *1994 International Symposium on Speech, Image Processing and Neural Networks*, pp.674-677., April 1994, HongKong.
- [7] F. Jelinek, Self-Organized Language Modeling for Speech Recognition *ICASSP'92* pp.450-506.
- [8] Y. Liu, N. Liang, Phrase Frequency Dictionary of Contemporary Chinese, *Yuhang Press*, 1989.
- [9] Idiom Dictionary of Contemporary Chinese, *Shangwu Press*, 1990.