REALTIME VITERBI SEARCHING FOR PRACTICAL
TELEPHONE SPEECH RECOGNITION SYSTEMS

Jin Zhang, Jia Liu, Run-sheng Liu
Department of Electronic Engineering,
Tsinghua University, Beijing 100084, China
zj98@mails.tsinghua.edu.cn

ABSTRACT
This paper studies searching and pruning process of the
telephone speech recognition system for Private Automatic
Branch Exchange (PABX) to explore the possible problems
encountered in applying speech recognition to telephone
network and to prepare the necessary techniques for the
practical telephone speech recognition systems. Experiment
on a baseline system which uses semi-syllable based multi-
subtree decoding structure and a classical Viterbi beam
search algorithm achieves 89.86% keyword accuracy rate. By
employing the dynamic threshold method, the keyword
accuracy can reach 93.48 %. By employing the ‘speed up
jumping strategy’, we achieve a higher performance
with 97.35 % in keyword accuracy.

1. INTRODUCTION
With voice controlled PABX, users can be connected to
adestination by speaking out a sentence including keywords like
a name, a site or a telephone code. The acceptability of the
service will be severely limited if users are told to change their
speaking style. To be user friendly, the voice controlled PABX
has to be accurate and robust. It should ignore both in-
vocabulary utterances that result in poor recognition score and
out of vocabulary (OOV) utterances.

Our work focuses on the issue of accuracy and allows users the
flexibility to speak naturally within the vocabulary of the task.
Here keyword spotting seems a ready candidate to fulfill the task.
However, although significant progress has been made in
keyword spotting, its performance degrades when vocabulary
size increases. Therefore, for a medium vocabulary application
like PABX, a more effective and simpler method is adopted, i.e.
detecting keywords by recording the recognition path in a multi-
subtree structured decoder[1].

This paper is organized as follows. Overview of the tree
structured recognition system is presented in section 2. Then
section 3 introduces the theory of beam search algorithm. Section
4 describes the paths protection methods including ‘stack-
structured preservation model’ and ‘speed up jumping strategy’.
Experimental results are discussed in section 5. And the paper is
concluded with section 6.

2. BASELINE SYSTEM

2.1 Acoustic Model
According to the characteristic of Chinese language, the semi-
syllable based Continuous Hidden Markov Model (CHMM) is
selected as the acoustic model of our telephone speech
recognition system. It is well known that Chinese is a
monosyllable-structured language. Each Chinese word (syllable)
can be divided into two semi-syllable parts: initial and final,
which corresponding to the consonant and the vowel of each
syllable. The HMM for each Initial and Final has 2 states and 4
states respectively with one full rank Gaussian distribution every
state. We can organize the semi-syllables into several lexical
prefix-subtree for the sharing of the same parts of word
sequences, and then use the hierarchical multi-subtree structure
to represent the whole search space[1].

2.2 Multi-subtree Structured Recognition
Through the grammatical analysis of sentences emerging from
general mandarin PABX, syntactic sentences are obtained. They
are commonly described as:

Figure 1. A Syntactic Sentence

Every syntactic sentence can be developed into a large number of
simple sentences when keywords are inserted. A rule sub-tree is
generated according to the structure of syntactic sentences.
Subtrees of site, name, title and code are set up respectively. The
basic multi-subtree framework is composed of the 5 trees above.
A decoding process is demonstrated in Figure 2.

We can detect Keywords expediently by recording the subtree
category and keywords number.

The baseline system achieves a keyword accuracy of 89.86% for
good Mandarin speech. Such performance is decent in view of
the simple acoustic models and high variability in speaking styles.
However, it is not satisfactory because false acceptance is very
costly for a PABX.
3. TIME-SYNCHRONOUS VITERBI BEAM SEARCH

In our telephone speech recognition system, a time-synchronous Viterbi beam search technique is developed as the real-time search algorithm. Under this search technique, all hypotheses are pursued in parallel. When a new frame of voices entered the decoder, the accumulative likelihood-scores of all active hypothetical paths are calculated. In order to conserve the computing and memory resources, it is imperative to identify low-scoring partial paths that have a very-low probability of getting any better, and stop them from propagating further. The process of removing such paths from the search space is known as pruning. By introducing a threshold of likelihood-scores, we can divide the active paths into two parts and discard all the paths whose likelihood-score are below the threshold. Pruning is an important way to maintain computational efficiency[2].

However, the pruning procedures may delete some partial paths that have correct phoneme sequence but temporary a low score. Thus will inevitably produce recognition errors and decrease the recognition accuracy.

4. PATHS PROTECTION METHODS

4.1 Concepts of Paths Protection

One of the most efficient ways on improving the recognition accuracy is to apply some a prior knowledge of the acoustic models and language models of the telephone speech before constructing the decoding network. In this section, we incorporate new paths protection methods into beam search algorithms to reduce the decoding errors, with low additional computation overheads. The opportunities of all partial paths to grow into correct recognition results are not equal. With some a prior knowledge, we pick out the partial paths that are most possible to contain correct sequence of semi-syllables, keep and develop them even though they have low likelihood-scores.

4.2 Stack-structured Preservation Model

Some partial paths that may have significant influence on the recognition accurate, or its current node is a pivotal node in the subtrees, or its current semi-syllable node has confusable pronunciation (in Chinese, such as /l/, /m/, /n/ in initial and /en/, /eng/, /ei/ in final), are picked out and marked with a protected flag. During the decoding progress a dynamic area of memory, called 'preservation stack', is located.
preservation stack. When it is reached, the protected paths that have lowest scores are discarded. Active partial paths without protected flags are directly discarded when their likelihood-scores are lower than the current pruning threshold of active paths.

Figure 5 Pruning Procedure Applying Preservation Term. As illustrated in Figure 5, for achieving more searching efficiency, we import a parameter called ‘preservation term’ to enhance the pruning method. When an active path is conveyed into the preservation stack, a deadline of time is assigned to it. Before the deadline, the path has an opportunity to become an active path over again if its likelihood-score increases above the threshold. If the preservation term expired and the path still has a score-score below the threshold, it will be deleted and stopped from propagating further.

4.3 Speed Up Jumping Strategy

When the last state of a semi-syllable has been loaded, a partial path is about to split into several subsequence paths. During the splitting period, a father partial path can produce a large amount of subsequence paths. Most of them has similar word sequences and should be merged to avoid overlapping. But the path merging always delays the state jumping time from last state of the current semi-syllable to first state of the next semi-syllable. If this kind of delay occurred continually, the corresponding paths will be difficult to reach the leaf node of the rule sub-tree and will not form a valid recognition result. To alleviate this problem, we can reserve several subsequence paths till at least one of them reaches the last state of the next semi-syllable.

5. EXPERIMENTS AND RESULTS

The baseline system uses semi-syllable based multi-subtree decoding structure and a traditional Viterbi beam search algorithm. It is trained with a task unrelated Chinese 863 Mandarin speech database including utterances of 200 speakers, 120 males and 80 females. The total units used are 143 context-dependent semi-syllables (100 Initials and 43 Finals). It is tested with a 40 speakers database, half each gender, consisting of 207 sentences commonly used in PABX applications. On the Pentium PC with 266MHz CPU and 64M memory, it has 89.86 % for keyword accuracy and 2.1 times of real-time (the duration time of sentences to be recognized) for average sentence recognition time.

Table 1 Test of Stack-structured Preservation Model

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Error Rate (%)</th>
<th>Rate Reducing (%)</th>
<th>Recognition Time (times real-time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>10.1</td>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td>+ Key Tree-Node Preservation</td>
<td>7.24</td>
<td>28.6</td>
<td>2.7</td>
</tr>
<tr>
<td>+ Confusable Pronunciation s Protection</td>
<td>6.52</td>
<td>35.8</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Table 1 shows recognition results on applying Stack-structured Preservation Model. First, we give partial paths protected flags only if the current nodes of them are important nodes on the semi-syllable based subtrees. The result shows that the relative error rate reducing are 28.6%. Second we add protected flags to those paths having confusable pronunciations current semi-syllables.

Table 2 Performance of Speed Up Jumping Strategy

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Error Rate (%)</th>
<th>Rate Reducing (%)</th>
<th>Recognition Time (times real-time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>10.1</td>
<td>0</td>
<td>2.1</td>
</tr>
<tr>
<td>+ Speed Up Jumping Strategy</td>
<td>2.65</td>
<td>73.9</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Table 2 shows the results of speed up jumping strategy. Compare with Table 1, the error rate are lower.

6. CONCLUSION

In this paper, we have developed new paths protection methods for beam search algorithms to improve the recognition accuracy. By employing the ‘Stack-structured Preservation Model’, we improve the recognition accuracy from 89.86% (10.1% error rate) to 93.48% (6.52 error rate), with an increase in average recognition time from 2.1 times real-time to 2.8 times real-time. By employing the ‘speed up jumping strategy’, we achieve a higher performance with 97.35 % in recognition accuracy and 3.3 times real-time in average recognition time. Since the average recognition time increases, some fast decoding algorithm ought to be investigated to maintain the computational efficiency.

7. REFERENCES