Study on the Strategy for Hierarchical Speech Recognition

Dali YANG, Mingxing XU and Wenhu WU

Center of Speech Technology,
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science & Technology,
Tsinghua University, Beijing, 100084
{ydl, xumx, wuwh}@sp.cs.tsinghua.edu.cn

ABSTRACT

A novel strategy for hierarchical speech recognition is proposed in this paper. Based on space partition, it takes the advantage of each recognizer on subspace. It organizes recognizers in a manner of nested recursion. Experiment results show that the final performance of the new method can reach an error reduction about 33% compared with the best recognizer.

Keywords: hierarchy recognition, multiple recognizers’ integration, mess degree, distribution entropy

1. INTRODUCTION

Hierarchical recognition method has been proposed for a long time in the pattern recognition society. Although output synthesis method has been widely adapted in other fields, it is not successful in speech recognition up to now.

When hierarchical recognition strategy is adapted in speech recognition, there are two difficulties: (1) the recognition frameworks are so similar and the types of feature parameters are too limited; (2) there isn’t any systematic and effective strategy on how to integrate the sub recognizers.

Usually, multiple recognizers’ integration in the speech recognition system is implemented as a integration of the output results. For example, Bayes method[2], Dempster-Shafer method [1][7], BKS method [4] and Sorting Combination methods [6]. Because the output results of multiple recognizers are used only once, it is impossible to take advantages of the space partition ability of each recognizer and to consider the different recognition performances on different unit subspace. Further more, although selection of recognizer is dynamic in some hierarchical multiple recognizers integration, the dynamic selection is processed only once[7]. In this paper a novel method is proposed to address the above problems.

This paper is organized as follows. In section 2 the principle and concepts of the new method is described. In section 3 the hierarchical structure building is given. In section 4 a hierarchical recognition method is presented. The experimental results are reported in section 5.

2. PRINCIPLE AND CONCEPTS

It is well known that the performance of the recognizer depends on unit space. When some unit spaces are given and they are able to sort in mean of set theory, then the smaller the set size is, the better the performance is. Hence, when a unit space could be divided into some smaller subspaces according to certain criterions, the final performance of the whole recognition system will be improved. It is obvious that the subspaces could be able to be divided continuously and recursively.

Let V stand for the unit space and M=(aij) stand for the mess matrix. There are two new concepts, i.e. mess degree (MD) and distribution entropy (DE), are introduced to evaluate the space division ability of the recognizer. They are defined as follows.

\[
MD = \sum_{j} N_j, \quad DE = \sum_{j} p_j \log(p_j)
\]

Here V={1,2,...,m}, Vj={i:aij>0}, Nj=Vj, N=MD, p=V/N. Vj is a set that includes all the units recognized as unit j.

MD is a measurement of the space division ability. The recognizer with a smaller MD will have the better performance. DE is a measurement of the uniformity of the subspace sizes. The recognizer with a bigger DE will have the better performance.

Let ER stand for the error rate. The recognizers are evaluated with the vector (MD, DE, ER) in selection process according to the following steps:

Step 1: r0=ArgMin{MD(r)}.

If r0 is unique, r0 is best recognizer, else go next step.

Step 2: r1=ArgMax{DE(r); MD(r)=MD(r0)}.

If r0 is unique, r1 is best recognizer, else go next step.

Step 3: r2=ArgMin{ER(r); MD(r)=MD(r0) and DE(r) = DE(r1)}.

r2 will be regarded as the final recognizer.

The evaluation process is called Space Partition Method (SPM).

3. HIERARCHICAL STRUCTURE BUILDING

In order to implement the hierarchical recognition, a tree is used to express the hierarchical recognition strategy, which is called Recognize Decision Tree (RDT).

Each node of RDT contains a vector (N, R, L), where N stands for the represent unit, R stands for the recognizer and L stands for the unit space.

The RDT building process is illustrated in figure 1, which includes the following steps:

...
Step 0: Let current node equals to root node of RDT and \( L \) equals to \( V \) and \( N \) equals to 1.

Step 1: According to the unit space \( L \) of the current node, subspaces set \( \{ V_j(R) : j=1,2,...,m, r=1,2,...,n \} \) is calculated.

Step 2: The evaluation vector set \( \{ (MD(r), DE(r), ER(r)) \} \) is calculated.

Step 3: Get recognizer \( R \) according to SPM described in section 2.

Step 4: When the unit space \( L \) is divided into several subspaces \( \{ V_j(R) \} \) with recognizer \( R \) obtained at step 3, its child-nodes are produced. The vectors set of the child-nodes is \( \{ (j, *, V_j(R)) \} \). The symbol "*" stands for that the recognizer \( R \) is not determined.

Step 5: For each child-node of the current node, repeat step 1 to step 4 until the following conditions (A or B) is satisfied.

Condition A: (Trivial Condition) The unit space \( L \) of the current node equals to the unit space \( L \) of the parent node.

Condition B: (Ideal Condition) The unit space \( L \) includes only one unit, that is \( |L| = 1 \).

4. RECOGNITION METHOD BASED ON RDT

The hierarchical recognition system is expressed as a tree (RDT) and its application is a trip of the RDT. Actually, the trip is from the root node of the RDT to some leaf node of the RDT. The detail steps are following:

Step 0: Let current node equals to the root node of the RDT.

Step 1: For a given sample to be recognized, the recognition result could be obtained with the unit space \( L \) and the recognizer \( R \) of the current node of the RDT. Let \( j \) note the result.

Step 2: Look for child-node \( c \), such that the unit \( N \) contained in the vector of the note \( c \) equals to \( j \). If the node \( c \) is found, let current node equals to \( c \) and then go to step 1, else unit \( j \) will be considered as the system’s recognize result.

5. EXPERIMENTAL RESULTS

5.1 Chinese digit recognition experiments

In order to get experiment results at lest time, we test the new method on a small speech data set that includes 10 digits 2839 sampling syllables.

5.1.1 Unit space

Ten Chinese isolate digits are selected as units, that is

\[ V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \]

5.1.2 Recognizers

Speech recognition software HTK 3.0 is used in our experiments. All acoustical models are based on HMM. Each unit has 6 states with one mixture. We selected 10 feature combinations. No.0 recognizer’s features are MFCC(13), D_A. No.1-No.9 recognizers’ features refer to our paper [8].

5.1.3 Speech data set

Speech data sampling rate equals to 16 kHz, 16bits/Sample. Speech data set is divided three sets: acoustic model training set (1772 syllables), regulation training set (710 syllables), testing set (357 syllables). Table 1 shows that result experiments on test set.

<table>
<thead>
<tr>
<th>Recognition Method</th>
<th>Recog. 0 (baseline)</th>
<th>Voting method</th>
<th>New Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Rate</td>
<td>97.48%</td>
<td>98.04%</td>
<td>98.32%</td>
</tr>
<tr>
<td>Error Rate</td>
<td>2.52%</td>
<td>1.96%</td>
<td>1.68%</td>
</tr>
<tr>
<td>Error Rate Reduction</td>
<td>22.2%</td>
<td>33.3%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: experiment results on test data

Note: No. 0 recognizer is the best one among 10 recognizers. Performance of the hierarchical method precedes the best with an error reduction of 33%. The method precedes Voting method.

5.2 863’s 100 syllables experiment

Recently, we had some results on national 863’s speech data. The experiment results refer to table 2 and table 3.

5.2.1 Unit space and speech data set

We selected 100 syllables data from 80 male speakers’ speech data. These syllables have enough samples for method testing. And the acoustical samples size notes as ASS, regulation sample size notes RSS and test samples size notes TSS. They satisfy that:

\[ \text{ASS} = 800, \]
\[ 200 \leq \text{RSS} \leq 400, \]
\[ 200 \leq \text{TSS} \leq 400. \]
5.2.2 Features of recognizers

8 recognizers’ features refer to table 2.

<table>
<thead>
<tr>
<th>Recognizer</th>
<th>dimension</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>126</td>
<td>LPCC(12) + MFCC[14] + E(2) + Var + PSC + LSP(12)_D_A.</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>LPCC(12) + MFCC(14)+ E(2) + Var + PSC + LSP(12)</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>LPCC(12)+ MFCC(14)</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>LPCC(12)+MFCC(14)_D_A</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>LPCC(12)+MFCC(14)+E(2) + PSC + Var</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>LSP(12)_D_A</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>MFCC(14)_D_A</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>LSP(12)_D_A</td>
</tr>
</tbody>
</table>

Here, MFCC stands for mel-frequency cepstral coefficients. LPCC stands for LPC-Cep. LSP stands for line spectrum pair. Symbol “_D” stands for delta coefficients. Symbol “_A” stands for acceleration coefficients. E(2) stands for absolute energy and absolute energy suppressed. PSC stands for power spectrum central. Var stands for power spectrum normal variance(refer to paper[8]). Each unit has 6 states with 8 mixtures.

Table 3 shows that hierarchical recognition strategy has no effect. Its performance is worse than voting method. On our opinion, the 8 recognizers is too familiar so that the new method no valid.

<table>
<thead>
<tr>
<th>Recognizer (Baseline)</th>
<th>Our method</th>
<th>Voting method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Rate (%)</td>
<td>87.14</td>
<td>87.42</td>
</tr>
<tr>
<td>Error Rate %</td>
<td>12.86</td>
<td>12.58</td>
</tr>
<tr>
<td>Error Rate reduction %</td>
<td>2.1</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Table 3: 100 syllables experiment results

5.3 Further work

We will continue study hierarchical recognition strategy and look for some new methods to use features that are able to improve MFCC’s recognition performance.

6. REFERENCES


