TOWARDS HIGH PERFORMANCE CONTINUOUS MANDARIN DIGIT STRING RECOGNITION

Yonggang DENG, Taiyi HUANG and Bo XU
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100080
Email: {ygdeng,xubo,huang}@nlpr.ia.ac.cn

ABSTRACT

In this paper, we address the problem of high performance speaker-independent continuous Mandarin digital string recognizer and focus on exploiting context information and prosody knowledge. Data-driven decision tree method to train tri-phone acoustic model was proposed. According to Chinese language property, digital specific question set was designed and the derived tri-phone model is more accurate to describe acoustic observation. For prosody cue, a novel Gaussian Mixture Density Duration Model (GMDDM) was presented. Unlike traditional normalizing or single parameter strategy, proposed duration model is context independent. The context variation is naturally embodied into multiple Gaussian distribution mixture. The number of mixture is automatically selected according maximum likelihood criteria. This simple but effective duration model’s likelihood score is combined with acoustic score as heuristic information for the backward A* decoding of word graph. Experimental results show the tri-phone acoustic model could lead to average 12.9% reduce of string error rate. When GMDDM model is applied, the string error rate is further reduced by 22.7%, which demonstrates the very usefulness of GMDDM model.

1.INTRODUCTION

Many researchers have investigated continuous digital string recognition. For English digital string, the performance of speaker-independent recognizer is good. There are acceptable applications like digital dialing and ID number voice input, even under telephone condition [4]. But in the case of Mandarin, there is obvious gap between the art of state and real word demanding. This is partly because Chinese is monosyllable and there are several highly confusing pairs, like “2” and “8”. Moreover, there are three digits pronounced as single vowels: “1”, “2”,and “5”. To segment vowel vowel sequence, such as “55”, is very difficult, which will result in high insertion and deletion error rates.

To obtain high performance speaker-independent continuous Mandarin digital string recognition system, we have employed a serial strategy from training to recognizing. Effectively capturing context information and prosody knowledge will make a remarkable improvement to recognizer. In this paper, we will only discuss decision tree based tri-phone acoustic model and duration model.

Acoustic model plays a very important role in small vocabulary speech recognition tasks like digital string recognition. Bahl [1] presented decision tree method. In [3], decision tree based context-dependent acoustic model was proposed, and in [2], tri-phone model was proposed with the tool of decision tree. Their studies show that decision tree based tri-phone model could significantly improve recognition accuracy and provide a pronunciation-optimization capability. To fully exploit the small vocabulary size, in our system, digit-specific question sets are designed for data-driven decision tree generation. We group the digit according to INITIAL or FINIAL part. Tri-phone models are derived automatically based on context-independent acoustic models driven by training data from top to down.

Prosody cue is an important knowledge source to disambiguate the word’s boundary. It has been widely accepted that segment duration carry useful and sometimes indispensable information for connected digit recognition. L.C.W. Pols et al [6] performed a detailed analysis of context-dependent duration variability, from the recognition points of view, for the TIMIT database. Their studies reached a discouraging conclusions that many duration features are not as consistent as expected and only some of them would be of potential importance for speech recognition purpose.

Two basic questions must be answered for each duration model strategy. The first is how to model the duration segment, and the second is how they can be incorporated into search process effectively. For the first question, there has been existing many duration model realizations in the literature [5][7]. Most of them are context-dependent, to be specific, duration variations is reduced by normalizing according to stress level, word position in utterance, speaking rate and etc. Existing methods can be observed like carefully organizing the duration information into nicely structured framework according to their context [5], or trying single parametric distribution such as Gaussian, Poisson or Gamma density [8]. As for the second question, the duration cue is either used to guide viterbi decoding [8], or re-score N-Best candidates in post processing [7].

In this paper, we escape the complex duration variability but propose a novel context independent Gaussian Mixture Density Duration Model (GMDDM). Unlike traditional ways, multiple Gaussian mixture is employed to capture duration variability. The number of mixture component for each word is variable...
and is obtained automatically based on Maximum Likelihood criteria. This simple but effective duration model’s likelihood is combined with acoustic score as heuristic information for the backward A* decoding of word graph, which is generated by time-synchronized forward viterbi beam search.

This paper is organized as follows. In section 2, we will describe the decision tree based training of tri-phone acoustic model. Duration model and its integration into decoding process will be described in section 3. Experimental results are presented in section 4 and conclusion is presented in section 5.

2. DECISION TREE BASED TRI-PHONE ACOUSTIC MODEL

To accurately describe acoustic variation, context information should be embodied into acoustic model. For connected digital string recognition, we only consider the left and right context, that is, tri-phone model.

For each model, observation vector marked with left and right context information was mapped to proper state by Viterbi alignment algorithm with context-independent acoustic model. Before constructing the tree, the following aspects should be considered: the question set, evaluation function and stopping criteria.

2.1 The Question Set

To fully exploit small vocabulary, we designed digital-specific question set. The left context is classified by FiNIAL part of Mandarin digit, while right context is grouped by FIRST part of the next digit. As table 1 shows, 10 groups for left context and 9 groups for right one were obtained. Every group corresponds to one question, resulting in 19 questions totally. In fact, for each side context, every model should be covered by one group, but is not limited to only one group. Each group is mapped to one kinds of context.

| Left context:          | {Silence}, {yao1}, {0}, {1,7}, {2}, {3}, {4}, {5}, {6,9}, {8} |
| Right Context:        | {Silence}, {0}, {1,2,yao1}, {3,4}, {5}, {6}, {7}, {8}, {9} |

Table 1: Digital-specific question sets

2.2 Evaluation Function

After one node is split, we need to know how good this splitting is. For each node, we will use a function to determine how similar those vectors are. Let \( Y \) represents the set of labeled data associated with node \( n \). \( M \) denotes the estimated probability distribution function (pdf) for the data set at node \( n \). For each pdf \( M_i \), \( p_M(y) \) describes the probability assigned to data \( y \). So \( P_M(Y_n) = \prod_{y \in Y_n} p_M(y) \) is a measure of how well the pdf \( M_i \) fits the data at node \( n \). If \( M_i \) is the best pdf for \( Y_n \), it describe the purity of the data at node \( n \), that is, if the data in \( Y_i \) are similar to each other, \( P_{M_i}(Y_n) \) will be large.

Suppose the node \( n \) is split by question \( q \) into two subsets \( Y_i \) and \( Y_r \). \( M_i \) and \( M_r \) denote to corresponding pdf, then the outcome of this split is

\[
m(n, q) = \frac{\log(P_M(Y_i)) + \log(P_M(Y_r))}{\log(P_M(Y_i))}
\]

Since \( m(n, q) \) is a measure of the improvement in purity as a result of this split by question \( q \), it is a very good choice for evaluation function. In our approach, we use the data at each node to obtain the pdf by k-means clustering, which is an appropriate estimation with not so expensive computational complexity. This evaluation function is advantageous to describe the global information between the model and the training data, and it is normalized to percent.

2.3 Stopping Criteria

If the outcome of the best question split at node \( n \) is less than threshold, or the number of the data at node \( n \) falls below threshold, we stop the split of this node and labeled it as a leaf node. The threshold for minimum evaluation increase is selected empirically, and the threshold for number of data is determined by the number of data in root node.

2.4 Constructing the Tree

The constructing procedure is from top to down, starting from the untested root node with all data belonging to it. Each time, we select an untested node, and try every untested question. If the outcome of the best one is more than threshold and both numbers of the data associated with each child node is more than threshold, we split this node by the best question and mark it tested, and its child node untested, otherwise we label this node as leaf node. This process is repeated until all nodes are tested.

For each base model, we build a decision tree for every state separately. After that, a merge procedure is operated to find all its tri-phones. Starting from the root, every kinds of context can reach a leaf node according to the split questions. For those context that reach the same leaf node for every state’s decision tree, we merge them into one tri-phone model.

3. GAUSSIAN MIXTURE DENSITY DURATION MODEL

Duration variability are too complex to capture. In this paper, we present an simple but effective Gaussian Mixture Density Duration Model(GMDDM). Like acoustic model, we use multiple Gaussian mixture to characterize the model duration distribution. And the mixture number is selected automatically by Maximum Likelihood criteria.

Suppose we have to estimate base model i’s duration distribution according to N duration data \( t_1 \| t_2 \| ... \| t_N \). Using k-Means clustering algorithm, we could figure out k mixture
duration pdf

\[ b_k(t) = \sum_{i=1}^{k} C_{\mathcal{N}(u_i, \delta_i^2)}. \]

And the mixture \( k \) is determined by:

\[ b_k^*(t) = \arg \max_{b_k(t)} \sum_{i=1}^{N} \log(P(t | b_k(t))) \quad (2) \]

GMDDM is context independent, thus the duration likelihood is easy to integrated into search process with little modification of decoding algorithm. In our recognizer, the duration score is linearly combined with acoustic likelihood as heuristic information for backward A* decoding operated on word graph, which is derived after forward time-synchronized Viterbi-Beam search.

Figure 1 Backward A* search on word graph

As figure 1 shows, when partial path \( P1 \) is ready to backward extended word \( w_i \), which spans from \( t1 \) to \( t2 \), the alpha function for new partial path \( P2 \) is defined as:

\[ \alpha(P_2) = \alpha(P_1) + \text{Acoustic\_Likelihood}(w_i) + \lambda \cdot \log(\text{Duration\_Score}(t1 - t2 | w_i)) \quad (3) \]

4. EXPERIMENTAL RESULTS

4.1 Database

Experimental database includes 36 female and 55 male data. Each person has 80 utterances. The vocabulary is “0” to “9” and “yiao1” and silence, totally 12 unites. After decision tree split, 466 tri-phone models was obtained for female data and 513 for male data. Although the mixture of context independent acoustic model is 16, which is accurate enough, and the mixture of tri-phone model is 4, we got about 80% increase of Gaussian distributions. It shows the tri-phones are more detailed, and the observation vector’s clustering is much more better.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Speaker of training</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td># of Speaker of testing</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Percent of Left Question</td>
<td>49%</td>
<td>47%</td>
</tr>
<tr>
<td>Percent of Right Question</td>
<td>51%</td>
<td>53%</td>
</tr>
<tr>
<td># of Tri-Phones</td>
<td>466</td>
<td>513</td>
</tr>
<tr>
<td>Uni-Phone Mixtures</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Tri-Phone Mixtures</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td># of Tri-Phone Gaussians</td>
<td>1972</td>
<td>1884</td>
</tr>
<tr>
<td># of Uni-Phone Gaussians</td>
<td>1072</td>
<td>1072</td>
</tr>
<tr>
<td>Increase of Gaussians</td>
<td>84%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Table 2 Comparing of Tri-Phone and Uni-Phone

As table 2 shows, right context has more effect (about 52%) on pronunciation, but its advantages is not obvious.

4.3 GMDDM and Tri-Phone Results

The baseline of our experimental system is two-pass search based on context independent continuous density HMM. The first pass is forward Viterbi-Beam search, which generates word graph. And the second pass is backward A* decoding.

As table 3 shows, when decision tree derived tri-phone acoustic model is used, the average string correct rate of testing samples is 91.39% and 90.50% for female and male respectively. Compared with baseline, the string error rate is reduced by 22.5% and 6.6% respectively. For training data, the Tri-Phone effect is more obvious. The average string error rate is reduced by 51.2% and 40.6% for female and male respectively. For all female and male testing set, the average string error rate is reduced by 12.9%.

<table>
<thead>
<tr>
<th>Search Methods</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Baseline</td>
<td>94.12%</td>
<td>94.16%</td>
</tr>
<tr>
<td>Baseline+Tri-Phone</td>
<td>97.13%</td>
<td>96.53%</td>
</tr>
<tr>
<td>Baseline+Tri-Phone+GMDDM</td>
<td>97.18%</td>
<td>97.38%</td>
</tr>
<tr>
<td>Tri-Phone effects</td>
<td>51.2%</td>
<td>40.6%</td>
</tr>
<tr>
<td>GMDDM effects</td>
<td>1.7%</td>
<td>24.5%</td>
</tr>
</tbody>
</table>

Table 3 Search Results for different methods
After GMDDM is applied, the string error rate is further reduced by 24.2% and 21.9% for testing set respectively, which resulted in the best string correct rate of 93.47% and 92.58% respectively. For all female and male testing set, the average string error rate is reduced by 22.7%. This is not conspicuous for training set.

Experimental results proof the efficacy of proposed tri-phone model and GMDDM duration model. It’s worth pointed out that the effects of tri-phone and GMDDM are complementary. Both will improve the performance of the recognizer independently.

4.4 Error Analysis
Most of recognition errors result from single vowels. It’s very difficult to figure out the boundary when the single vowel digit “1”, “2” and “5” is juxtaposed with other digit. For example, “71” and “7”, “65” and “6”, “52” and “5”. They lead to very high insertion error and deletion error. To prevent such kinds of wrong recognition, prosody information, especially duration, is indispensable. Together with detailed context dependent acoustic model, effective duration model play magnificent help in the improvement of recognizer’s performance. Only about 18% errors come from replacing, which could be reduced by discriminative training.

5. CONCLUSION
Continuous Mandarin digital string recognition is more difficult than English language. We discuss the problem of the high performance connected digital recognizer. A serial strategy for more accurate model and better decoder have been adopted. In this paper, we only discuss context information and prosody knowledge.

Data-driven decision tree method to obtain tri-phone acoustic model was presented. Digital specific question set was designed. Derived tri-phones models are more detailed, and the observation vector’s clustering is much more better.

A novel Gaussian Mixture Density Duration Model (GMDDM) was proposed. This model is context independent, and the context information is embodied naturally into multiple Gaussian distribution mixture, like traditional acoustic model.

Experimental results show that, when the tri-phone is used, the average string error rate of recognizer was reduced by 22.5% and 6.6% for female and male test set respectively. When the GMDDM model is applied, the performance of recognition is even better, with average string error rate reduced by 24.3% and 21.9% for female and male test set respectively. And the tri-phone and duration model improved the performance of recognizer independently.

Acknowledgements
This work is partly funded by Nokia China. The authors wish to thank helpful discussions with Mr. Yuan Dong and their colleagues.

REFERENCES