On the Application of Associative Method to Chinese Connected Digit Recognition

Wang Jiu-Long, Yuan Xue-Gong, Zhao Guo-Tian

Speech Signal Processing Lab.
Department of Radio Engineering
Harbin Institute of Technology
Harbin, P.R. China

ABSTRACT

In this paper, a recognition algorithm based on the characteristics of Chinese is proposed for the real-time recognition of Chinese connected digit using an associative method.

A Chinese character is a CV monosyllable. A Chinese digit string is easy to be segmented into individual digits which carry little information. The present DTW method only takes account of entire difference between two digits regardless of local difference whose existence sometimes is enough for us to distinguish the two digits from each other. We have found:

1. the recognition performance will improve when recognition results from two separated parts of one digit are associated.
2. the recognition performance will improve further when recognition results from different distance measures are associated.

Using above recognition method, we have conducted several experiments on four-to-five-digit strings. A 16-ch BPF is used as spectral analyzer. For the speech spoken at normal speed (about 200 to 250 digits/min.), the recognition accuracy is greater than 97% with a response time less than 0.5 second.

I. INTRODUCTION

Recently, many algorithms have been proposed to perform the recognition of connected digits spoken in English and other. Each of these algorithms has used either of two recognition ways. One way is to segment the spoken digit string into single digits and then recognize each digit respectively [1]. The other way is to recognize digit string directly using DTW method without pre-segmentation [2].

According to the characteristic of Chinese digits that segmentation of digit string is easy to realize, the first recognition way is applied for our purpose. Then the question lies in the recognition of single digits segmented. As known, one Chinese character is a monosyllable which carries little information, the difference between different digits sometimes is small and the discreteness of one digit in different digit strings is large because of coarticulation. All these disadvantages make the current DTW method by itself difficult to act well.

In this paper, an associative algorithm considering local information of one digit and making use of advantages of two kinds of distance measure is presented which succeeded in the real-time recognition of Chinese connected digit.
II. RECOGNITION PRINCIPLE

A. Associative Decision on Two Local Candidates

From the table I we can find that the phonematic difference between some two digits, 6 and 9 for instance, is very small. Human can differentiate 6 from 9 if he had perceived the difference between phoneme [l] and [j] regardless of the diphthong [iu]. But when applying DTW method, all local distances are averaged to obtain an accumulated distance, once the local difference, between [l] and [j] for example, is small enough not to have a significant influence on the accumulated distance, then error recognition is possible. To improve the recognition accuracy one should not allow the local information submerged in the course of average.

Based on the point of view mentioned above, a natural approach is to segment the whole digit into two parts: C and V, then to recognize each phoneme respectively. In fact, the difficulty of inter-phonemic segmentation is not less than recognition itself, so that a simple effective method be required. On our purpose, each digit is segmented into approximately equal two parts, the spot at which the digit segregates is decided by DTW algorithm. Thus each digit have two recognition results: head candidate and terminal candidate. The association of two candidates can yield a more reliable result.

Assume $h(n,k), k = 1, 2, ..., Kn$ represent possible head candidates of digit $n$ ($Kn$ is the number of candidates), and $t(n,k), k = 1, 2, ..., Mn$, represent possible terminal candidates of digit $n$ ($Mn$ is the number of candidates). So we have gotten a head associative file $H(n) = (h(n,k), k = 1, 2, ..., Kn)$ and a terminal associative file $T(n) = (t(n,k), k = 1, 2, ..., Mn)$. Thus each digit $n$ has three relevant files: one reference template and two associative files. The two associative files have been compiled up depending on phonetic knowledge prior to recognition. In DP-matching if the test digit is the same as reference digit, generally the head recognition result includes in $H(n)$, and the terminal recognition result includes in $T(n)$. But if the test digit is different from reference digit, at least either of the two candidates do not include in relative set $H(n)$ or $T(n)$. Assume $h(k)$ represents the head candidate of test digit $k$ and $t(k)$ the terminal candidate, if $h(k) \in H(n)$ then should $t(k) \in T(n)$, vice versa. Otherwise the test digit $k$ would be recognized as digit $n$. Take test digit 6 and reference digit 9 for example. $H(9) = (3, 4, 7, 9)$, $T(9) = (9, 6, 5)$. It is very likely that $t(6) \in T(9)$ because digit 6, 9 and 5 have identical terminal phoneme [iu], but the head recognition result should’t be one of 3, 4, 7, 9, for phoneme [l] in digit 6 has a great difference with the fricative of digit 3, 4, 7, 9. Therefore, usually digit 6 can’t be recognized as 9. By the same reason digit 9 couldn’t be recognized as 6.

B. Associative Decision on Recognition Results From Different Distance Measures

<table>
<thead>
<tr>
<th>Phonetic Symbol of Ten Chinese Digits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ling</td>
</tr>
<tr>
<td>1 yi</td>
</tr>
<tr>
<td>2 liang</td>
</tr>
<tr>
<td>3 san</td>
</tr>
<tr>
<td>4 si</td>
</tr>
</tbody>
</table>
Our experiments show: using different distance measures, different recognition results can be obtained. Of these results any one could be incorrect, but their association sometimes benefits for the recognition.

Assume \( T = (t_1, t_2, \ldots, t_n) \) is a frame of test speech features, \( R = (r_1, r_2, \ldots, r_n) \) is a frame of reference speech features. Follow simple local distance measure is referred to as absolute distance:

\[
d(T, R) = |T - R|
\]  
(1)

This distance measure is related to the energy of \( T \) and \( R \). Another distance measure disregards the energy information of speech. We can derive follow distance formula (called relative distance) from the cosine of angle of vector \( T \) and \( R \):

\[
d(T, R) = 1 - \frac{(t_1 r_1 + t_2 r_2 + \cdots + t_n r_n)}{(t_1^2 + t_2^2 + \cdots + t_n^2)^{1/2} (r_1^2 + r_2^2 + \cdots + r_n^2)^{1/2}}
\]  
(2)

Consider that above local distance is not easy to compute, we have derived follow formula from equation (2):

\[
d(T, R) = \frac{1}{2} |T/|T| - R/|R||^2
\]

Replace above equation with:

\[
d(T, R) = |T' - R'|
\]  
(3)

Here, \( T' \) and \( R' \) are test and reference speech features which are all normalized on energy.

The two distance measures have their advantages and disadvantages. Generally, consonant in head of a digit is short in length and weak in energy, but it plays an important part in speech recognition. The relative distance can eliminate the disadvantage of small energy by discarding energy information but it exerts simultaneously a big weight on noise fields and causes extreme speech feature distortion. On the contrary the absolute distance is not sensitive to noise fields or feature discreteness of speech intervals with small energy, so show better recognition performance for some digits such as 0, 1, 2.

C. Realization of Algorithm

Although the segmentation of digit string is successful based on pattern recognition principles [1], end ambiguity is inevitable. To eliminate the effect of segmentation impreciseness, a kind of fixed-window Level-Building DTW algorithm is used to time-align the concatenated digits. Because of successful pre-segmentation the region of digit boundary is confined extremely. Hence, the region in which all local distances have to be computed is far smaller than that of current LB-DTW algorithm [2].

Fig.11 shows the principle of matching algorithm. Before DP-matching, each digit is time-aligned linearly to the same length as reference template. Assume input test digit string consists of \( l \) digits, then \( 2l \) level LB-DTW is to be computed. To get two recognition candi-
dates, each digit needs a two-level LB-DTW computation.

In matching at some levels, the vocabulary could be less than ten. For digit $n$, assume the head candidate $h(n)$ has been obtained, we can get a subset of ten digits:

$\{ k | h(n) \in H(1_k), k=1,2,\ldots,K \}$

Therefore the terminal matching is restricted on above subset.

III. EXPERIMENTAL RESULTS

In order to estimate our algorithm proposed above, several experiments on Chinese four-to-five-digit string have been carried out. A 16-ch BPF is used to get speech spectral feature parameters. The speakers are four males, with normal speaking rate (about 200-250 digits/min.). All programs had run on an IBM-PC/XT microcomputer without any fast-processing part.

At first input digit string is segmented into single digits according to the position of one digit in the string, three sets of reference template are used, clustered respectively from first digits, last digits and embedded digits in all reference digit strings. Each digit is only to match with the corresponding reference template.

We had taken 100 strings to train the reference templates. Another 1000 strings were adopted for test strings. The experimental results are:

1. If taking whole digit as a DP-matching unit, the recognition accuracy is poor (less than 90%)
2. The recognition accuracy applying associative algorithm using one distance measure is about 95.5%.
3. The recognition accuracy of associative algorithm using two kinds of distance measure is about 97.3% and the recognition time is about 0.5 second.

IV. CONCLUSIONS

Finally, we conclude that:

1. The associative method presented in this paper is simple and effective to Chinese connected digit recognition. It shows that making use of the characteristics of Chinese is beneficial for the recognition of Chinese speech.
2. If taking account of contributions of all local informations and all parameter forms of speech feature to the recognition of the word, the recognition accuracy can improve distinctly.

REFERENCES