ABSTRACT
This paper presents a new context dependent tone recognition method. First we suggest that there be more than five tone modes in Chinese continuous speech. We get all new tone modes by grouping all tone feature vectors to a specific number of categories. Secondly, we recognize a sentence with the new tone modes and get the new tone sequence. Finally, we find out each original tone of the sentence which has the maximum a posteriori probability for the corresponding new tone and its context new tones. In a ten-person test set, which includes about 20,000 tones, we achieve a higher average recognition rate with the MAP-based context dependent tone method than that with the conventional context dependent tone method.

1. INTRODUCTION

The Initials, Finals and tones are the three essential factors in Chinese phonology. Tone information is a crucial source for correctly understanding the meaning of a sentence. There are many Chinese words that have the same pronunciation without considering tones. It will be very difficult to catch the meaning of a sentence if there is no tone information in it. Therefore, the tone recognition plays an important role in Chinese speech recognition. Many efforts have been made in this field. [1][2][3][4]

The isolated syllable tones recognition is comparative easy in tone recognition. When a person speaks isolated syllables, the pronunciation is often replete. Thus, the correct rate of tones is above ninety percent in general. [1] But the tone recognition in continuous speech is much more complex and difficult. In the continuous speech the duration of some utterances is so short that we can not extract the fundamental frequency contour. The most serious situation is that people frequently change from an original tone mode to another tone mode in continuous speech especially in the third tone mode, which increases the similarity of different tone modes and gives rise to a sharp decline in recognition rate. Considering above reasons, we put forward a MAP-based context dependent tone recognition approach.

Conventionally people think there are five tone modes that include four basic tone modes and a neutral tone mode. But in fact there seems to be more than five tone modes in continuous speech. We call them the new tone modes. Of course there must be some kind of relationship between the original tone modes and the new tone modes. Then, how can we obtain the new tone modes? The resolving way is quite simple.[5] We can get the new tone modes by classifying all tone feature vectors of training data into several categories using the classical clustering algorithm.[8] Each central vector of every category corresponds a new tone mode. Using the new tone modes to recognize a sentence, we can attain a sequence of new tones. Because context tones interact greatly [7], we have to take it into account. Intuitively, people will apply the context dependent tone models to recognize tones.[6] But because there are too many parameters that we have to train for the context dependent tone models, it may lead to decreasing the correct rate for the insufficiency of train data. Therefore, we propose a MAP-based context dependent tone recognition method which we only need much less models. When we get a sequence of new tones, we choose the original tone on each position that has the biggest probability on the condition of the current new tone, preceding new tone and successive new tone.

Section II will describe the details about the training and recognition method of MAP-based context dependent tone. The experiment result will be shown in section III. The summary is finally given in section IV.

2. THE METHOD OF TRAINING AND RECOGNITION

2.1 Obtain the New Tone Modes by Clustering Algorithm

Let P denotes a sequence of raw fundamental frequency for one syllable. Since the length of the sequence is always changing. In order to take advantage of clustering algorithm, P has to be normalized using the following formula. [5]

\[ P(i) = \sum_{j=\text{floor}((i+1)K/D+1),k=1}^{\text{ceil}(K/D),k=k+1} P(j) \times W(k) \]

Where, K is the original sequence length and D is the desired length. \( \text{Ceil}(x) \) rounds the elements of x to the nearest integers towards infinity, \( \text{Floor}(x) \) rounds the elements of x to the nearest integers towards minus infinity. \( W(k) \) is a triangle window.

The average fundamental frequency (F0) of every person is different and it may vary even for the same person in different sentences. Our goal is to build up a speaker independent tone recognition approach. So we have to normalize tone feature vectors before training and recognition. We add up all elements of all tone feature vectors in a sentence and divide it by the number of elements of all vectors in the sentence. In this way we get the average F0 of the current sentence. Then, the every element of each vector in the sentence is divided by the average F0. After that, the tone feature vectors that have been normalized are obtained.

Then, we put all tone feature vectors of all persons together and classify them to K centers, which represent K new tone modes respectively.
2.2 Train the Context Dependent Probability Matrix

We recognize a sentence in the train set using the new tone modes upon the nearest mode rule. By this way we attain a new tone sequence. On the other hand, the original tone sequence is known. There must be some kind of relation between the two sequences. In the continuous speech the context tones have effect on each other. In order to model the coarticulation effects suitably, we define a context dependent matrix:

\[
C = \{ c_{i,j,k,m} \mid \Pr(R_i = o_m \mid S_{i-1} = v_i, S_i = v_j, S_{i+1} = v_k) \} \\
1 \leq i,j,k \leq K \\
1 \leq m \leq M
\]

where,

- \( R \) represents the sequence of original tone
- \( S \) represents the sequence of new tone
- \( t_R \) and \( t_S \) represent the \( t_{th} \) original tone and the \( t_{th} \) new tone respectively.

\( O = \{ o_1, o_2, \ldots, o_M \} \) is the set of original tone modes.

\( M \) equals to 5.

\( V = \{ v_1, v_2, \ldots, v_K \} \) is the set of new tone modes. It is suitable that \( K \) is between 8 to 20.

\( \Pr(x) \) represents the probability.

The meaning of the above definition is that \( c_{i,j,k,m} \) is the probability of that the \( t_{th} \) original tone is \( o_m \) on the condition of that the current new tone is \( v_j \), the preceding new tone is \( v_i \) and the successive new tone is \( v_k \).

Obviously, the definition shows the interaction of context tones. We can obtain the C matrix through the following formula:

\[
c_{i,j,k,m} = \frac{\sum_{M=1}^L \sum_{i=1}^L \gamma(i,j,k,m)}{\sum_{M=1}^L \sum_{i=1}^L \sum_{m=1}^M \gamma(i,j,k,m)} \\
1 \leq i,j,k \leq K \\
1 \leq m \leq M
\]

Where,

\[
\gamma(i,j,k,m) = \begin{cases} 
1 & \text{if } S_{i-1} = v_i, S_i = v_j, S_{i+1} = v_k \text{ and } R_i = o_m \\
0 & \text{otherwise}
\end{cases}
\]

Because the pause on the beginning and end of sentence also has effect on the context tones, we regard the beginning/end as a kind of new tone \( S_{pause} \). In this way we resolve the problem that the new tones at the beginning of sentences have no preceding tones and the new tones at the end of sentences have no successive tones.

While extracting the pitch contour, some duration of utterance is so short that we can not obtain their F0 sequences. In this case we regard the tones that can not attain tone feature vectors as a new tone mode \( S_{failed} \). By this way we can still give the candidates upon the MAP rule even without the valid tone feature vectors.

2.3 Recognize Tones

Taking advantage of the new tone modes to recognize a sentence, we can get a sequence of new tones \( S_1, S_2, \ldots, S_r \). After inserting the beginning tone and the end tone, the sequence will be \( S_{pause}S_1S_2\cdots S_rS_{pause} \). If there is a tone \( S_r \) which is failed to get its pitch contour, then let \( S_r = S_{failed} \). Applying the following formula to the sequence, we can attain the original tone sequence.

\[
R^*_r = \arg \max_{o_m \in O} \Pr(R_r = o_m \mid S_{r-1} = v_i, S_i = v_j, S_{r+1} = v_k) \quad 1 \leq i \leq T
\]

Where,

\[
R^* \text{ is the optimized sequence of original tones.}
\]

\[
\Pr(R_r = o_m \mid S_{r-1} = v_i, S_i = v_j, S_{r+1} = v_k) = c_{i,j,k,m}
\]

3. EXPERIMENT RESULT

In order to compare with the traditional method of context independent tone (CIT)\[1\] and context dependent tone (CD)\[6\], we also give the result of CI and CD. For the consistency of modeling tone method, we apply the classical clustering algorithm\[8\] instead of those complicated algorithms such as HMM. Following is the brief description of CI and CD.

1) CI models

Put all the tone feature vectors with the same tone mode together and get all CI models by clustering. Each tone mode may have several sub-tone modes.

2) CD models

Put together all the tone feature vectors that have the same current tone and the same context tones and then obtain all CD models by clustering. Every tone mode may also have several sub-tone modes. While recognizing tones, the Dynamic Programming is applied.

For the reliability of the result, plenty of train data and test data are used. The train set includes about 100,000 tones of fifty speakers. The test set includes approximate 20,000 tones of ten speakers.

The distribution of the each tone mode in the test set is below:

<table>
<thead>
<tr>
<th>Tone Mode</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>5.6%</td>
<td>22.7%</td>
<td>21.8%</td>
<td>16.4%</td>
<td>33.5%</td>
</tr>
</tbody>
</table>

Table1: The distribution of each tone mode

<table>
<thead>
<tr>
<th>Tone Mode</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top One</td>
<td>36.5%</td>
<td>69.5%</td>
<td>51.9%</td>
<td>26.5%</td>
<td>47.5%</td>
</tr>
<tr>
<td>Top Two</td>
<td>66.0%</td>
<td>81.3%</td>
<td>68.1%</td>
<td>61.1%</td>
<td>63.4%</td>
</tr>
</tbody>
</table>

Table2: The recognition rate with the CI tone method

<table>
<thead>
<tr>
<th>Tone Mode</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top One</td>
<td>26.6%</td>
<td>64.0%</td>
<td>55.4%</td>
<td>44.7%</td>
<td>59.4%</td>
</tr>
<tr>
<td>Top Two</td>
<td>74.8%</td>
<td>75.0%</td>
<td>72.1%</td>
<td>81.8%</td>
<td>68.1%</td>
</tr>
</tbody>
</table>

Table3: The recognition rate with the CD tone method
Tone Mode | 0 | 1 | 2 | 3 | 4  
--- | --- | --- | --- | --- | ---  
Top One | 5.6% | 65.8% | 61.1% | 38.4% | 70.8%  
Top Two | 20.1% | 79.7% | 77.9% | 69.0% | 88.2%  

Table 4: The recognition rate with the MAP-based CD tone method

The following figure is the average correct rate of top one and top two with the method of CI, CD and MAP-based CD.

![Figure 1: The average recognition rate of top one and top two using the method of CI, CD and MAP-based CD.](image)

4. CONCLUSION

For the purpose of suitably dealing with tone coarticulation in the continuous speech, we put forward a MAP-based context dependent tone recognition method. It avoids the drawback in the conventional CD tone method that there are so many parameters in the CD models that may degrade performance. In the case that the pitch contour can not be extracted, we can still give the valid candidates, which partly solves this problem. Finally, the experiment result shows the new method is effective and the average recognition rate with MAP-based CD tone is higher than that with the conventional CD tone.

5. REFERENCES