ABSTRACT

Tonal Evaluation of Chinese continuous speech plays an important role in Mandarin Chinese Pronunciation Test. In this paper, we introduce the Multi-Space Distribution Hidden Markov Model based on prosodic word. The results show that the performance of tonal syllable error rate can be reduced. For the non-standard Chinese Mandarin speech, the correlation between computer score and expert score was improved above 3.0% absolutely, compared with the baseline system without tonal pronunciation test.

Index Terms—Mandarin Chinese Pronunciation Test; Tonal Evaluation; Tonal recognition; MSD; Prosodic word; Mandarin speech recognition.

1. INTRODUCTION

One of the central functions of Computer-assisted Language Learning is to assess the pronunciation level and detect the pronunciation defects by computer. Tone is an important scale which distinguishes dialect from Mandarin, and it can even determine a person's Mandarin level. Therefore, in the whole Mandarin standard test system, the objective tonal evaluation is an important subsystem. The technology of tone recognition is the basis of tone assessment. Only by the accurate tonal recognizer can we get the reliable and valid result of tonal assessment system, so the first step is to get a recognition system which can achieve a high recognition rate on the standard pronunciation database. It is not difficult to recognize tone on isolated word, however, when faced to the continuous speech in which the tone curve is badly influenced by context, there is continuity that can't be ignored between the tones of syllable and the various existence of tonal modification, which causes confusion to the results of tone recognizer, so it will seriously influence the actual performance of tonal evaluation system.

To deal with variations in fundamental frequency (F0) contours for tonal recognition in continuous Chinese speech, Jingsong Zhang divides F0 contour of a syllable into tone nucleus and adjacent articulatory transitions[1][2]. Tone nucleus of a syllable is assumed to be the target F0 of the associated lexical tone, and usually conforms more likely to the standard tone pattern than the articulatory transitions. Guokang Fu presents experimental results on Cantonese and Mandarin[3]. While the number of phonemes and the quantity of training data are substantially reduced, the decoding accuracy is improved over other methods by using the main vowel method. However, in many cases, the accurate location determination of tone nuclei as one essential prerequisite for the proposed method is difficult, and integrality of the core segment and rationality of the extending compensation part cannot be guaranteed by the existing methods. On the usage of the feature continuity, it is common to set up the contextual model based on the tone nuclei segment extraction, but in fact this expansion has not fully utilized the continuity among words.

How to utilize tone co-articulations and how to solve the F0 contour discontinuity between voiced and unvoiced segments are the two problems we discuss in this paper. The approach of building the MSD-HMM based on prosodic word is proposed here[4], and the recognition results show that the performance of tonal syllable error rate can be improved.

In Section 2, the principle and composition of tonal evaluation of continuous speech based on prosody information are described. The results and analysis of the experiments are presented in Section 3. The conclusion is drawn in Section 4 and followed by a prospect for the future.

2. SYSTEM COMPONENTS

In this paper we take the tonal recognition technology as the basis of tonal evaluation, so we need a tonal recognition system which can give good performance on the standard pronunciation database, that is, correct pronunciation database, to make sure the technology used to the tonal level test is suitable and believable. Thereby, the system introduced here is composed of two segments: Tonal recognition on the standard speech database and tonal evaluation on dialect speech database.

The system include 5 modules mainly: pitch extraction, pitch procession, text analysis, model training, tonal
recognition/tonal evaluation. The schematic diagram is as follows:

![Diagram of Tonal Evaluation System](image)

Firstly, by the pitch extraction module the accurate and robust fundamental frequency will be supposed to be got from continuous speech, however, there always be some mistakes like half-frequency and double-frequency. The influence of context may cause variations of fundamental frequency (F0) contours in the continuous speech, that is, the difference of F0 contours between the continuous speech and the isolated units is great. So the target of second module is to give comparatively accurate F0 features after a series of rectifies. By the text analysis module, we can get the prosodic structure label in order to give the boundary of prosodic words. Then the system trains the 16MIX MSD-HMM model by using of tool HTS-2.0. Finally, adopting the method of recognition on restrained boundaries, the module of tonal recognition and tonal evaluation give the result as the posterior probability and evaluation score separately.

2.1. Pitch Procession

It is not difficult to conclude that, the pitch feature of speech is decided mainly by following factors: 1. Pure tone of isolated syllable. 2. Some statistical fluctuations and changes. 3. Context of the speech. 4. Pitch range of the speaker. The first 2 factors will be easily described, while the other 2 factors lead to the instability and variations of tone. To deal with the problem, we need some suitable rectifies. Firstly, 16mix HMM model trained on the 863 standard continuous mandarin speech database is used to calculate the boundary of initials and finals. Then, the tone nuclei segment will be extracted. We calculate the difference from the middle sample to the both ends separately, and set the point which has a difference bigger than thresh as starting point or finishing point. Besides, there will be a mean normalization system to lower the impact of speakers. Finally, we make use of the technology of long-term pitch normalization (LPN) to reduce the context influence.

2.2. Text Analysis

The training and testing text can be analyzed directly because of the peculiarity of speech evaluation system itself. The main step is text pretreatment, words segmentation, post-treatment, initials-finals classification. Then by use of the prosodic structure analysis, we can get the prosodic boundary for tone model training.

In the general tonal recognition system, because testing text does not exist, we can classify the border of initials and finals (candidate border) according to its acoustic characteristic in order to produce the prosodic words boundary label. The acoustic characteristics are mainly shown as follows: Science lies between the borders; Extension of the syllable before the borders; Fundamental frequency is reset.

2.3. MSD-HMM based on Prosody Information

2.3.1. Prosodic Words

In communications, the prosody information plays a very important role in the naturalness and Intelligibility of speech. The various levels of continuity provided by prosody information split continuous speech into smaller units in order to make it easy to understand and deal with by machine, and to provide important basis for eliminating syntax ambiguities.

Modern Chinese phonetics shows that prosody is closely related to changes of fundamental frequency contour. Therefore, the tone information can be inspected according to the frequency changes of prosodic structure theoretically. The feature used for the recognition system presented here includes two aspects, that is, fundamental frequency feature and the corresponding prosody structure. Prosody structure describes the information of different types of interval information in the pronunciation according to grammar and semantics, or even pronunciation habits. This information can reflect some regularity of the variations of fundamental frequency in continuous speech.

Prosodic word plays a key role in bigger prosodic composition. The resets of F0 between prosodic words and the characterized regularities of F0 within the prosodic words are the important characteristic of spoken language. Prosodic word, as the group unit of F0 changes, has no perceptible pause existing between them. The interplay effect is much more serious in the prosodic words than high-level prosodic events. The usage of prosodic word unit, take replacement of traditional contextual unit, and setting up models distinguishing the circumstances of boundary tone from others, it can better simulate the real pronunciation and reduce the confusion of the model.

We will show the important effect of tone on the prosodic word structure through the following example: “从前面对面”， we can find two kinds of structure: “从前/面对面” and “从/前面/对”。The corresponding pitch curves are compared and analyzed in fig2.1 and fig2.2.
The beginning pitch point of “面” (TONE4) does not clime high enough in figure 2.1 because the “面” is spoken after the character “前” of TONE2, and that two characters constitute one prosodic word. In figure 2.2, when “面” lies in a prosodic word which is composed of the character “前” and “对”, the end part of “面” does not dropped to the normal tone pitch as the influence of the following character “对” of TONE4. So the conclusion is that the restriction within the prosodic word is obviously exceeds that between prosodic words.

As the prosodic word unit is composed of one or several syllables, How to deal with the discontinuity of F0 between the voiced and unvoiced segments cannot be neglected here. According to the principle of MSD-HMM, as to fundamental frequency, we set up two probability distribution space, dispersed and continuous, correspondent fundamental frequency, we set up two probability ingredients, How to deal with the discontinuity of F0 between the voiced and unvoiced segments cannot be neglected here.

3. Experimental Results and Analysis

3.1. Experimental Setup

Database used in this section is recorded from PSC test, which is a national test to evaluate the proficiency of spoken Mandarin. It contains four parts of test. Our tone error detection experiment is carried out on the 3rd part of PSC test, the reading text consists of 60 articles, which have 400 words in average. Training set is collected as 28 people (standard mandarin data including 15 female speakers and 13 male speakers).

Experiment 1: Tonal recognition on standard set. Testing set consists of 9 people (standard mandarin data including 4 female speakers and 5 male speakers). The experiment adopts standard pronunciation data (pronunciation level recognized as first level) as test database to tonal recognition. The higher recognition rate of standard data, the more reliable evaluation system is. Experiment 1 include three sub-experiments as follows: Tonal recognition based on textual tonal model(Tritone), tonal recognition based on prosodic word adopting the interpolation method(L1Tone), tonal recognition based on prosodic word adopting MSD-HMM(L1Tone-MSD).

Sub-experiment 1: Feature Procession, training contextual HMM, 180 model units.

Sub-experiment 2: Feature Procession, linear interpolation of the unvoiced part inner prosodic word according to text analysis. Training HMM model based on the prosodic word, with 123 model units. The model unit consists of no more than 3 characters, with bigger units splitting into smaller ones.

Sub-experiment 3: Feature Procession, training 16MIX MSD-HMM model based on prosodic word by HTS-2.0, with 123 model units.


\[
b(o) = \sum_{g \in \Omega(o)} p(\Omega_g) p_x(o)
\]
province, 100 people with 100 articles separately. The baseline system is mandarin pronunciation test system according to the spectrum feature of the words, phrases, chapter and the F0 feature of the words and phrases based on dialect data from Henan and Shandong province. The purpose of this experiment is to analyze the promotion of system performance by adding the tonal evaluation sub-system of continuous mandarin speech based on the MSD-HMM of prosodic word unit.

3.2. Result Analysis

Experiment 1:

<table>
<thead>
<tr>
<th></th>
<th>Tritone</th>
<th>LiTone</th>
<th>LiTone-MSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>82.0%</td>
<td>83.1%</td>
<td>84.6%</td>
</tr>
</tbody>
</table>

Table 1. Performance Comparison of Tritone\LiTone\LiTone-MSD system

<table>
<thead>
<tr>
<th>TONE</th>
<th>TONE1</th>
<th>TONE2</th>
<th>TONE3</th>
<th>TONE4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>90.3%</td>
<td>82.6%</td>
<td>84.2%</td>
<td>81.0%</td>
</tr>
</tbody>
</table>

Table 2. Results of tone recognition of LiTone-MSD

Table 1 compares the results of 3 sub-experiments. Here, TONE0 is out of the statistic because of its complicated situation. It can be found that, the performance of LiTone-MSD is the optimum, LiTone takes second place, Tritone is the worst from the table. In experiment 1, Tritone unit of prosodic word edge is different from that of inner prosodic words, the extension of contextual tone model on the edge of prosodic words is unreasonable. Therefore, when the two situations are mixed to train the same Tritone model unit, the obscurity of the model may increase. In the second sub-experiment, we introduce the method of interpolation which is an artificial assumption and lack of authenticity to LiTone, the performance is not good. LiTone-MSD of the third experiment gives the best result.

Experiment 2:

Table 3. Performance Comparison of Baseline and LiTone-MSD system on dialect database

<table>
<thead>
<tr>
<th>CC</th>
<th>Shandong</th>
<th>Henan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.1%</td>
<td>54.7%</td>
</tr>
<tr>
<td>LiTone-MSD</td>
<td>69.5%</td>
<td>58.5%</td>
</tr>
</tbody>
</table>

As the data tested in experiment 2 has serious pronunciation problem, we can find obvious promotion by the addition of tonal evaluation system over baseline. The Cross-Correlation (CC) between the experts scores and computer scores is improved by 3.4% and 3.8% on Shandong and Henan database respectively.

4. CONCLUSIONS

This paper introduces the MSD-HMM based on prosodic word and proves performance improvement in the tonal recognition and tonal evaluation experiments. Prosodic structure information makes good use of continuous information between words and increases the differentiation between tones. In future work, we will try to introduce more prosody information to enhance the evaluation system performance.

5. REFERENCES


