1. INTRODUCTION

With the improvement of the computing power and the memory capacity of computer, large speech corpus can be made for the research in phonetics and prosody now. On the other hand, the synthesized speech by segment concatenation is effective and intelligible. So, the segment concatenation method based on corpus is a good way to produce speech with high naturalness. The main two problems of this method are: First, what content should the corpus include? Second, how to select the synthesis units in the corpus for a given target sentence? The two issues relate with each other closely.

With an infinite corpus, which can contain any utterances, there is no problem to produce the same speech as natural speech. But we only have finite corpus now. To get natural synthesized speech, the content of corpus should cover as many phonetic and prosodic conditions as possible. In Chinese there are only over 400 syllables without tone, and over 1300 syllables with tone. It’s a universal way using syllables with tone as the elemental units in Chinese speech synthesis systems [1][2]. All these syllables in different context will be different in prosody character, so the corpus designed should contain as much syllable variations as it can.

Which unit in the corpus is the best candidate for the target syllable to be synthesized? It is another problem to be solved. We believe and prove that, the different instances of a syllable may have different prosodic parameters such as pitch, duration and power and etc., they can substitute each other, and no perception difference is arisen unless the speaker is intent to change his speak style. Based on this idea, we can select the unit according to its context. This method needs no prosodic parameter prediction. There is also no signal processing need to modify the prosodic parameters that can degrade the speech quality.

2. CORPUS PREPARATION

2.1 Corpus Design

As said above, the more phonetic and prosodic cases the corpus contains, the better synthesized speech quality is. The elemental units in our system are syllables with tone. The contexts we consider include: the position of a syllable in prosody word (PW), the position of the prosody word in prosody phrase (PP), the position of the prosody phrase in breath group (BG), the tone of the previous and next syllable. The People Daily is used as source text. It is changed to phonetic string. The homographs in the sentence are manually checked. The phonetic string is scored according to the new context information referred above in it. When its score is above the threshold, the corresponding sentence is selected with its previous and next sentences into the text corpus to be recorded, forming a short paragraph, and the syllable context information database is modified according to the new context in phonetic string. The block diagrams of corpus design is showed on Fig.1. The threshold and the termination criteria are manually set. The final text corpus is read by a female speaker with broadcast experience in her normal speaking style. The recorded signals are sampled at 16kHz and quantitized at 16bit. So, a female speech corpus of 137M bytes is finished.
2.2 Segmentation and Labeling

The speech in corpus should be segmented and labeled for use. Segmentation is used to decide the position of each syllable. Labeling is used to determine the prosody structure. The prosody structure is divided into three levels from lower to upper: PW, PP, and BG.

The Segmentation of corpus is realized by HTK. The HTK units are right-context INITIAL and context-independent FINAL. The feature vectors include: 12 dimensions of MFCC, first and second order delta MFCC, RMS power. The result of segmentation is the position of INITIAL and FINAL of each syllable in speech. Some of them have to be manually checked, especially in the case that a series of No-INITIAL syllables adjoin in the speech.

The labeling of corpus is manually finished. The prosody structure is set at the boundary of syllables. An utterance may contain one to several BGs, and the judging standard of a BG position is the long pause between two syllables. While a PP position is judged by the short pause, F0 reset, or syllable lengthening of its ending syllable. The PW position is judged by perception mainly because there is no clear clue in it, and it has correlation with the utterance content. These are just the reason why we have to label by hand.

2.3 Index Building

After the segmentation and labeling, the F0 parameter is gotten by the software of ESPS. For each syllable, an index is built to get its information quickly. The index includes the following content:

1. Position information (beginning, ending, and INITIAL-FINAL boundary position in speech file);
2. Prosody parameters information (F0 value of beginning, ending, maximum and minimum);
3. Prosodic structure information (the name of current, preceding and next syllable with tone, the position of current syllable in PW, the position of the PW in the PP, the position of the PP in the BG). All these positions are classified into four types: head, end, middle and single.

There are also two F0 value in the index, the ending F0 value of the preceding syllable in the same PW (PrePitchEnd) and the beginning F0 value of the next syllable in the same PW (NexPitchBeg). All the F0 values in the index are normalized according to the speaker’s pitch range (60Hz-380Hz):

$$F_{0,norm} = 16 \times (\ln F0_{max} - \ln F0_{min}) / (\ln F0_{max} - \ln F0_{min})$$

3 UNIT SELECTION

3.1 Two Costs

The input of synthesis module is phonetic string with prosodic labeling; the output is synthesized speech. In [4], two costs used in the unit selection in corpus-based speech synthesis are presented: target cost $C(t_{ij}, t_j)$, and concatenation cost $C_{ij}(u_{i-1}, u_i)$. They are defined as following equations.

$$C(t_{ij}, t_j) = \sum_{j=1}^{p} w_j C_j(t_{ij}, u_j)$$
$$C_{ij}(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j C_{ij}(u_{i-1}, u_j)$$

The target cost describes the distance between the target unit and the instance unit in corpus, which is the weighted result of each sub-feature distance of the two units. The concatenation cost describes the distance between two consecutive instances. It is also the weighted result of each sub-feature distance of two consecutive units. The units to be selected finally should meet the requirement that the total cost of all these units is minimized.
This idea is not only used in CHATR system [5], but also used in Chinese speech synthesis [1]. The difference lies in the definition of two costs. These two systems both need the prosodic parameters (pitch, duration and power) prediction. Two problems are arisen. First, the prediction maybe is inaccurate; Second, maybe there is no unit in the corpus matching the prediction, and signal processing methods has to be used to modify the unit parameters. Both these probabilities will degrade the synthesis speech quality.

In our system, according to the context of target, the instance with the same syllable name and tone as the target in the corpus is selected whose context match the target’s best. So, there is no prosody parameter prediction, and no prosody parameter modification needed. The output is quite natural. According to the two costs referred above, we divide the unit selection process into two stages: candidate selection and path selection.

3.1 Candidate Selection

In this stage, for each target syllable, we select no more than ten candidates in the corpus to be used in the next stage.

According to our phonetic knowledge [2], four context-matching items are set: position of the syllable in PW, position of the PW in PP, position of the PP in BG, the pre-tone and next-tone of the syllable. Each instance with the same syllable name and tone as the target unit in the corpus will be checked. For each context-matching item, if the instance isn’t the same as that of the target, a target sub-score will be set to this instance, and the sum of the four sub-scores is the target cost of the instance. The sub-scores are defined manually. Ten instances with the minimal target cost are selected as the candidates for the use of next stage.

If there is no instance in corpus with the same syllable ID as the target, an INITIAL-FINAL connection method is used to produce the syllable. The units containing the initial parts or final parts are selected as described above. The pitch difference in connection position should be the lowest. The method is showed in Fig. 2.

3.2 Path Selection

The candidates selected in the last stage form a big lattice, including several small lattices whose target syllables belong to the same PW. The consecutive cost of two consecutive candidates in the same small lattice is calculated by the PrePitchEnd and NexPitchBeg in their index. The method is showed in Fig.3.

Two costs are used to calculate the concatenation cost: forward cost and backward cost, which are defined as the following equations:

\[
Forward\ Cost = \text{Abs}(\text{NexPitchBeg}_{i+1} - \text{PitchBeg}_i) \\
Backward\ Cost = \text{Abs}(\text{PrePitchEnd}_i - \text{PitchEnd}_{i+1})
\]

The concatenation cost between unit I and I-1 is calculated as the sum of these two costs. Function Abs()
is used to get the absolute value. Using the Viterbi algorithm, we can get the final units’ path in each small lattice. All these paths of the small lattices form the selected path of the whole large lattice. It is showed in Fig. 4. The number in bracket is the target cost of each candidate.

4. EVALUATION AND DISCUSSION

An informal compare in naturalness of synthesized speech was made between this corpus-base system and KD-863 system [2]. KD-863 is a TTS system based on the perception quantification technology. In 1998, a formal evaluation of speech quality for the Chinese TTS systems was held by the steering committee of the State High Technology Development Project of China, in which KD-863 got the top score in naturalness.

In our informal evaluation, three paragraphs were given. Each had about 300 syllables and the content dealt with essay, sports report and political editorial. Five students who could speak standard Mandarin were chosen to evaluate these systems by synthesized sentences. Each sentence in paragraph was played many times to avoid the influence of the play order on perception. The final test result of one system was given by the following equation.

\[ GR = \frac{\sum_{i=1}^{5} G_i}{(5 \times N)} \]

Where \( N \) is the total number of sentences in a certain paragraph, while \( G_i \) stands for the number of the sentences whose synthesized speech is preferred by the \( i \)th listener. The result is shown in the table 1.

<table>
<thead>
<tr>
<th></th>
<th>Essay</th>
<th>Sports report</th>
<th>Political editorial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus-based System</td>
<td>80%</td>
<td>95%</td>
<td>100%</td>
</tr>
<tr>
<td>KD-863 System</td>
<td>20%</td>
<td>5%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 1: comparison between two System

Judging from the above result of our informal test, we can draw such a conclusion that the Corpus-based TTS is better than KD-863 in naturalness of synthesized speech. The system has better naturalness in the synthesis of political paragraphs than that of essay and sports report.

There is much manual intervention in the corpus design, such as segmentation and prosody labeling. Although we checked the corpus carefully, there are still inevitable errors that may degrade the output speech quality. The checking work is also time-consuming and tiresome. At present, it will take a week to prepare a same size corpus from design to producing speech. One of our future researches is to automatize the whole process, but more knowledge on phonetics and prosody will be needed.

In the stage of candidate selection, the target sub-costs are set manually according to the knowledge on phonetics. How to integrate these costs with perception is another problem to be solved in future.

At present, the concatenation cost is calculated in the same prosody word. There is no attention on the cost of the consecutive prosody words and consecutive prosody phrase. Sometimes, there are some perceivable leaps between prosody words. To get more natural speech, appropriate cost of prosody words and prosody phrase should be added in the future.

5. REFERENCES