A COMPARISON BETWEEN SYNTHETIC SPEECH AND NATURAL SPEECH OF CHINESE

LU Shinan  HE Lin  YU Ge  FENG Yongqiang
Institute of Acoustics, Academia Sinica, Beijing
lunslinfo.unet.net.cn

Liu Juan
City University of Hong Kong, Hong Kong
b5420189@plink.cityu.edu.hk

ABSTRACT

In this paper 15 synthetic Chinese sentences provided by four typical Chinese TTS system have been analyzed, and compared with natural speech. Results reveal the remarkable differences between natural speech and synthetic speech including the temporal organization and intonation, which are the essential cause of degrading naturalness of synthetic speech. Therefore the parser and prosody design are emphasized for developing a new Chinese TTS system.

1. INTRODUCTION

With an effort in the past 15 years, the Chinese TTS has started in the application in some communication system. When it is applied to a simple system with small vocabulary, such as a computer weather forecast system, Chinese TTS realized by concatenation of non-uniform units is with high naturalness. But for unlimited vocabulary Chinese synthesis up to today the improving on naturalness is still a great challenge.

Since 1994 national evaluation of speech synthesis is carried out every two years supported by national developing high technology project. But there are only few tests for prosody of synthetic Chinese in the national evaluation. Here more acoustic analysis for the synthetic speech from last national evaluation (1998) has been done in order to supply a gap of the results of national evaluation.

The definition of “naturalness of synthetic speech” is not explicit yet. In the gross it is the difference between synthesis speech and natural speech in perception. We like to know what differences are there between natural speech and synthetic speech, or while the synthetic speech sounds not natural for Chinese?

2. SPEECH MATERIAL

Radical speech material are 15 sub-sentences selected from a text, the essay “中国人爱喝茶”(Chinese love tea)”, for the naturalness evaluation of synthetic speech, in which the syllable “茶/Cha2/” is included in each sentence as shown in followings:

1. 中国人爱喝茶1
2. 南京人自然也爱喝2
3. 老年人喝3用小紫砂壶，
4. 咏起词来慢慢几地品茶4味儿。
5. 他们选品5茶的最佳境界谓之 “回甘”。
6. 遂也深遂喝茶6。
7. 茶7叶种类粗算起来分为红茶8、绿茶9、花茶10。
8. 解先生说在家喝11和上茶12馆喝茶13味儿又不同，
9. 在家喝不如在茶14馆喝味儿厚。
10. 后来上茶15馆看看，
11. 才觉那里喝茶16与家里果然不同。
12. 曾老说他年轻时便爱上茶馆茶17馆儿。
13. 进茶18馆时是为了喝茶19，
14. 而在茶20座几上坐下，
15. 喝茶21的心情就变为香热间了。

The speech materials of synthetic speech of four typical Chinese TTS system: Ss1, Ss2, Ss3 and Ss4 are taken from the digital record of evaluation work field; of the corresponding natural speech are supplied later. The natural speech by Liu and Chen were recorded in City University of Hong Kong; others by Bxf, Hy, Jmj and Jy were recorded in Beijing broadcast institute. The speakers are young student, whose daily speech is Chinese.

3. SPEECH ANALYSIS

Speech materials are analyzed by use of multi-Speech Model 3700. The waveform, pitch contour and spectrum of selected sentences are analyzed. The main results are described as following:

The syllable duration

There are 15 sentences or 121 syllables in selected material. The durations of each syllable have been measured. In a general way the duration depends on the syllable structure (initial, final and tone), and the pre-stored syllables of natural speech are used as a synthetic unit, so the general distributions of syllable duration of synthetic speech are similar with natural speech as shown in Fig. 1. If the effect of syllable structure is removed, or seeing about the duration variety for same syllable, the difference is immediately appeared. The duration of 茶
Fig. 1 The duration distribution syllable 
/cha2/(tea) in above 15 selected sentences for 4 synthetic speech and 6 natural speeches are measured as shown in table 1.

Table 1 The duration of syllable /cha2/ in selected sentences

<table>
<thead>
<tr>
<th>No</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>Liu</th>
<th>Chen</th>
<th>Bxf</th>
<th>Hy</th>
<th>Jmj</th>
<th>Jy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>3</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>4</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>5</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>6</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>7</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>8</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>9</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>10</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>11</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>12</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>13</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>14</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>15</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>16</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>17</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>18</td>
<td>280</td>
<td>296</td>
<td>321</td>
<td>350</td>
<td>280</td>
<td>300</td>
<td>330</td>
<td>260</td>
<td>290</td>
</tr>
<tr>
<td>19</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>20</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
<tr>
<td>21</td>
<td>300</td>
<td>331</td>
<td>360</td>
<td>390</td>
<td>321</td>
<td>350</td>
<td>380</td>
<td>310</td>
<td>340</td>
</tr>
</tbody>
</table>

In this case the duration of syllable /cha2/ will be affected by the context, according to CAO’s(1) standpoint, for example the No.1 in Table 1 is at front of the sub-sentence boundary, but the No.2 is that of the sentence boundary, so No.1 will be more extended than No.2. And according to the syntax of sentence 7, the prognosticating duration of three /cha2/ are arranged as No.7 < No.10 < No.8 and No.9. It is basically truer for natural speech, but for synthetic speech some time is truer, some time is not. The Pearson correlation among different speeches is shown in Table 2.

Table 2 The Pearson correlations among different speeches

<table>
<thead>
<tr>
<th></th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>Liu</th>
<th>Chen</th>
<th>Bxf</th>
<th>Hy</th>
<th>Jmj</th>
<th>Jy</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>.07</td>
<td>.23</td>
<td>.30</td>
<td>.26</td>
<td>.24</td>
<td>.24</td>
<td>.39</td>
<td>.11</td>
<td>.28</td>
</tr>
<tr>
<td>S2</td>
<td>-.10</td>
<td>.52</td>
<td>.56</td>
<td>.20</td>
<td>.47</td>
<td>.66</td>
<td>.33</td>
<td>.55</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>.22</td>
<td>-.13</td>
<td>-.28</td>
<td>-.08</td>
<td>-.07</td>
<td>.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>.60</td>
<td>.12</td>
<td>.44</td>
<td>.67</td>
<td>.26</td>
<td>.68</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level
** Correlation is significant at the 0.01 level

It shows the great difference between synthetic speech and natural speech. For natural speech, expect speaker Chen with Jmj and Jy the correlations are significant above the 0.05 level. It means although each speaker have own speak manner, the universal duration rules are existent. For synthetic speech the correlations are very different, because each system has different method of duration control. The durations of syllable /cha2/ synthesized from Ss1 are not correlative with any one of others. Acoustic analysis shows that, the Ss1 only two pre-stored syllables /cha2/ with 300 and 280ms duration used to synthesize speech. Therefore it is hard to suit the natural speech. The measured duration in Table 1 for the Ss3 shows they are modified, but the negative correlation means the duration rule is inaccuracy. The better correlation of synthetic speeches by Ss2 and Ss4 with natural speech means they have a proper duration modification.

The temporal organization in sentence

As an example, the waveforms of synthetic Chinese sentences (No. 8) are arranged on the top in fig. 2, which are provided respectively by four typical Chinese TTS system; one of natural speeches lies on bottom. It shows the remarkable difference between synthetic speech and natural speech in time domain.

For the natural speech, the sentence is divided into four
prosodic phrases, “解先生说”，“在家喝茶”，“和上茶馆喝茶” and “味儿又不同” by the long pause of 300 ms, 240 ms and 280 ms respectively in the time domain. Although the subject(解先生) and the verb of predicate(说) are different components in syntax, if the subject is not long, the subject will be with the verb to form a prosodic phrase. On the whole the prosodic structure is consists with syntax shown in Fig.3.

解先生 说 在家喝茶 和上茶馆儿喝茶 味道又不同

Fig. 3 The syntax of sentence No. 8

Unfortunately we do not find out the phrase structure in the temporal organization of synthetic speech. There are no evident silences among phrase boundaries, which are shown by bestriding red lines in Fig.2. The Ss1 and Ss2 do not set long silence for the phrase boundary; the Ss3 and Ss4 set long silence, but not at the right position.

The intonation

The corresponding pitch contour of waveform in Fig.2 is shown in Fig. 4. The remarkable differences are also presented in the frequency domain. For natural speech if one pays attention to the low point of pitch as shown by red line, the declination will be find out in the prosodic phrase, and it will be reset to up at the begin of following phrase. And there is a prominence in prosodic phrase, at which the high pitch is presented (corresponding to the syllable or word of thick Chinese characters). Therefore some syllables in sentence are accented. According to SHEN's[3] standpoint the top line of Chinese pitch is modified by accent and intonation of sentence, while the base line is modified by rhythm. The pitch contour of natural speech in Fig.4 is typical. The alternant of pitch accent and un-accent, and with the temporal organization together form the rhythm. As a result, the semantic structure presented by voices is clear in natural speech. The natural rhythm not only gives us enjoyment, also to understand speech easily.

For synthetic speech we could not find out the pitch declination and resetting in Fig. 4. Expect Ss1 the pitch ranges of sentence 8 for different speeches are close to each other, which is measured in octave and shown in Table 3. Although they have enough pitch range to modify the stress, but the pitch movement is stochastic and disordered.

<table>
<thead>
<tr>
<th>Ss1</th>
<th>Ss2</th>
<th>Ss3</th>
<th>Ss4</th>
<th>Liu</th>
<th>Jy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>1.40</td>
<td>1.67</td>
<td>1.74</td>
<td>1.77</td>
<td>1.78</td>
</tr>
</tbody>
</table>

The Ss1 uses a handle of reducing pitch rang, in order to escape the mistaking of setting prosodic phrase and accent, but the synthetic speech sounds tedium and aridity.

4. DISCUSSION

Chinese parser

All evaluated Chinese TTS system run short of parser. At present the syntax of input text is one and only practical clue for designing the prosody of synthesis speech. The absence of syntax of converted text is the crucial cause of incorrect prosody of synthetic speech. As a result there are the remarkable differences between synthetic speech and natural speech in the temporal organization and intonation. The Chinese parser should be the one of main modules in next generation Chinese TTS.

Prosodic control

We measured the duration, pitch, intensity and spectrum of selected sentences for synthetic speeches, as well as natural speeches. The difference may be concluded in each acoustical aspects. In this paper we only point the difference of prosody, in particular the duration and pitch, because it is the key to improve the naturalness of synthetic Chinese. Prosodic characteristics of synthetic speech should be the keystone of next national evaluation of speech synthesis. The innovative evaluation method should be advised, which is in favor of detecting insufficiency and improving naturalness of TTS system.

Synthesize from word by word to phrase by phrase

We pay great attention to the Chinese prosody control of phrase level. Review the history of Chinese TTS development, before 10 years when Chinese syllables had been clearly synthesized, some one thought the Chinese unlimited vocabulary synthesis has been finished. But virtually the public does not accept the synthetic speech spoken syllable by syllable. Although the synthetic speech spoken word by word is more natural than syllable by syllable, the main rhythm block of Chinese is prosodic phrase, not prosodic word. Currently we are looking forward to a synthetic speech spoken phrase by phrase.

Neutral intonation

This term of “neutral intonation” means the basic intonation, which not or less including the effect of emotion, for example, in which only the grammatical accent is considered. At present it is a right strategy, the prosody design of Chinese TTS should be based on neutral intonation. ZHEN etc.[3] validated grammatical accent rules by psychological experiment. But call a handle, which only reduces pitch range of Chinese syllable, but does not modify them, as “neutral intonation” is wrong.
Data or rule driving

Data driving technique is successfully used not only in speech recognition and also in speech synthesis. As demonstrated in 1999 by INTER, for limited vocabulary the synthesis is very successful, but for unlimited synthesis, the problem pointed as above still exist. Although in data driving system the function of rule is indirect, the rule is effective yet. Based on brainpower of human the rule is able to discover by fewer data. We hope the cooperative research of psychology, phonetics and acoustics will enrich knowledge of Chinese prosody, and perfect the rule system in not far future.

5. CONCLUSION

From limited materials the results of acoustic analysis reveal the remarkable differences between natural speech and synthetic speech. We believe the differences including the temporal organization and intonation are substantial for affecting the naturalness of synthetic speech. The research of Chinese prosody is emphasized in order to perfect the prosody rule. The Chinese parser and prosody design are suggested as the main modules for next generation Chinese TTS.

ACKNOWLEDGMENT

Many thanks are due to the Department of Chinese, Translation and Linguistics, City University of Hong Kong who kindly provided the support of this research, and to Prof. Eric Zee for his help and many advices.

REFERENCE


[2] 沈炯 (1999), 汉语音高载流系统模型，《中国语言学的新拓展》石锋、潘悟云主编，香港城市大学出版社。

Fig. 2 The waveform of synthetic speech and natural speech of selected sentence No.8

Fig. 4 The pitch contour of synthetic speech and natural speech of selected sentence No.8