IMPLEMENTATION TO ASSESSMENT OF SPEECH SYNTHESIS SYSTEMS FOR CHINESE ON NETWORK

Ge Yu, Jialu Zhang, Shiwei Dong
Institute of Acoustics, Academia Sinica
Beijing 100080, China
Email: yu@west.ioa.ac.cn
Zjl@west.ioa.ac.cn

ABSTRACT

Under the support of Intelligent Computing Program, National 863 Project, a national assessment of speech synthesis systems for Chinese has been carried out since 1994. On March 1998, the 3rd testing was carried out in Beijing, and four different systems were evaluated. Both phonetic (acoustic) modules and linguistic modules of speech synthesis and the ability of text pre-processing of six TTS systems were examined. All of the testing materials for the TTS systems were distributed and the outputs were gathered through the network. We will introduce the test methods on the network. The testing results of speech Articulation and naturalness in MOS (Mean Opinion Score) were given.

INTRODUCTION

From 1994, there are three assessments up to now. There is much difference between the first and the second assessments in the contents of the tests. But the tests' methods of these two are not very individually. The second time, we had more experience, more systematic, more rigorous, smoother progress, but it had no much difference in the distribution of the tests' contents and the gather of the results and so on. We had thought over the shortcoming of the two assessments, and gathered the suggestion and opinion of the other experts. In the assessment of 1997, we designed a whole new test method and carried out. The automatization of the test contents' distribution, the gathering of the results, the sequence arrangement and the speech replay was realized.

TESTING MATERIALS

1. Speech intelligibility tests

1.1. Syllable lists KXY1-10

- Phonetically balanced, 10 basic lists
- 75 syllables/list divided into 25 trisyllable groups
- Carrying phrase: 第X组是XXX. (The Xth group is XXX)

1.2. Word lists KXC 1-N

- Phonetically balanced
- 100 words/list divided into 25 four-word groups
- Each group is a Semantically Unpredicted Sentence (SUS)

1.3. Sentence lists KXJ 1-N

- Simple sentences
- No more than 7 words in a sentence

2. Modified Rhyme Tests

- MRT lists
  21 syllables with different initial consonants and one syllable with no initial consonant.
- Five candidates of which the initial consonants have similar manner of articulation but different place from that of tested syllable.
- Carrying phrase: 我读X字 (I read X character)

3. Total quality evaluation

3.1. Testing materials

- Different styles (total ~ 2000 characters)
  (1) Business page, (2) Literary works, (3) Political eassy,
  (4) Sports section, (5) Science section
- Linguistic compact
  (1) Homographs (Personal names, place

names),

(2) Digital strings (years, telephone numbers, and numerals)

(3) Symbols and metrological units

(4) Special tone modification rules

3.2. Testing methods

- 5 point Mean Opinion Scale (MOS) with plus(+) and minus(-) marks
- 5 cycles of presentation in random order for tested systems.
- 20 listeners (16 common people, 4 experts in speech synthesis)

4. Text processing ability tests

- Expansion of the testing materials for total quality evaluation (4000--5000 characters)
- Linguistic compact (total testing points ~200)
- Both text files and waveform files should be sent back to testing center.
- Automatic correction

5. Anti-interference ability tests

- Speech intelligibility tests under different S/N ratios, S/N > 30 dB, S/N =15dB, S/N = 5dB
- White noise was added to synthetic speech through Audio console.

NETWORK STRUCTURE AND NAMING SYSTEM

1.The network structure:

At the first and second assessments, after we distributed the testing materials, the attended system began to synthesize the speech. And then, the attended systems replay the speech waveform themselves. Thus, we can not control the replay time accurately. To save time, each attended system will replay all of the outputs at one time. So the interval of the replay of the same materials of the different system will be long time.

From the first and second assessments, we know that the replay time and the different replay sequence will influence the testing results.

To resolve these problems, networking assessment should be a good choice.

This time, we adopt 10BASE-T ether net. The network connection protocol is NETBEUI and Microsoft network users. To thought abort the low system security required, the network server had used the Microsoft Windows 95 operating system, and the attended systems should connected the server by the corresponded protocol. The network group's name is TTSGROUP. The attended system has its own sharing directory in the network server. The directory name should be "S01, S02, …"(means system No.1, system No.2, etc). And when the tests were carrying out, each attended system could get the test contents from the sharing directory, and after the deal with, put the result to the same directory.

Because the attended systems are not quite a few (just four systems), and the quantity of data was not very big, so the bandwidth of 10 Mbps was good enough this time.

The network topology was shown in graphics 1.

2. Naming system

The testing materials were given by the pure text formatted files. For all of the testing materials, we asked the attended system to give us two results: the Chinese Pinyin files and the synthesis speech waveform files.

From the first and second assessments, we got one experience that the replay time and the different replay sequence will influence the testing result. In order to control the replay time and rearrange the replay sequence, all of the testing materials were divided into pieces this time. The small segments can let us control them easily.

All the pieces of the materials were put into directory in the server for the different attended systems. All the outputs produced by the attended systems were going to store in the same directory. In the server's sharing directory include the test materials and the results at same time. In fact, there are many small files in the directory (will be thousands). We have a naming system to divide them. The file name is composed by 6 department, "S" represent "system", "xx" represent the system number, "N" represent the different materials, "zz" represent the sequence of the materials, "EXT" represent the different outputs of the materials. The detail as the Table1:

Table 1. The testing materials and outputs files naming system.

SxxNy-zz.EXT	
S	Always be "S"
Xx	Digital number
N	"c" represent word
	"y" represent syllable
	"j" represent sentence
	"z" represent total quality evaluation
	"w" represent text precessing
	"m" represent MRT
Y	Digital number
Zz	Digital number
EXT	"txt" represent original test contents
	"py" the Chinese Pinyin
	"wav" the synthesis system's speech waveform.

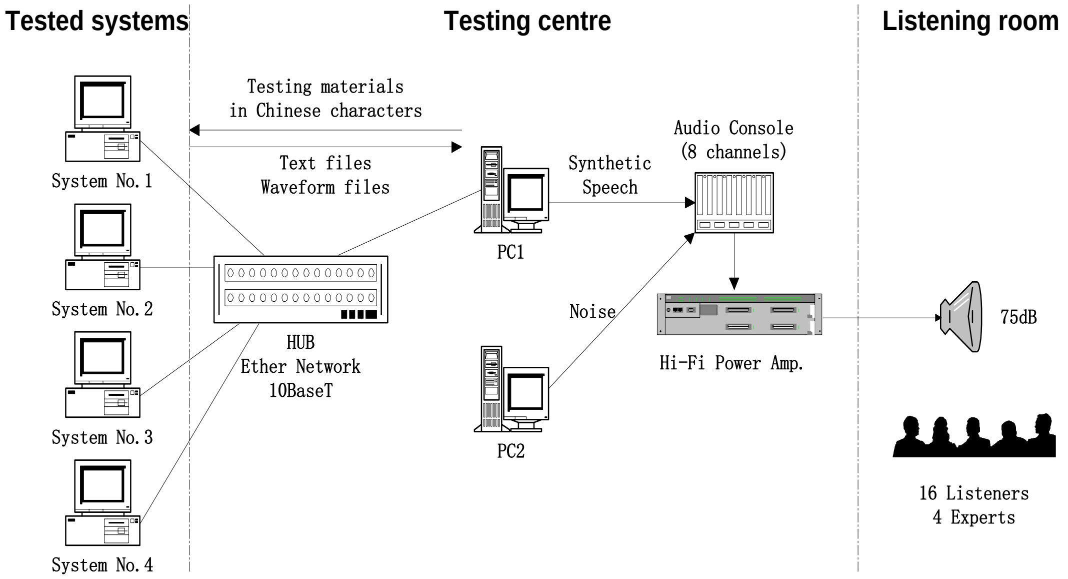
THE TESTING PROCESS

1. The attended systems ballot for the testing sequence.
2. Distribute the testing materials.
3. Attended systems make the outputs (to make the Pinyin files and to synthesize the speech waveform).
4. Gather the outputs from the attended systems.
5. Rearrange the output files as the sortilege.
6. Replay the waveform.

RESULTS

The testing results of speech Articulation and naturalness in MOS (Mean Opinion Score) were shown in table 2 , chart 1 and chart 2. Both the average value of intelligibility Av and its standard deviation σ were given for each system.

Graphics 1. The network topology.



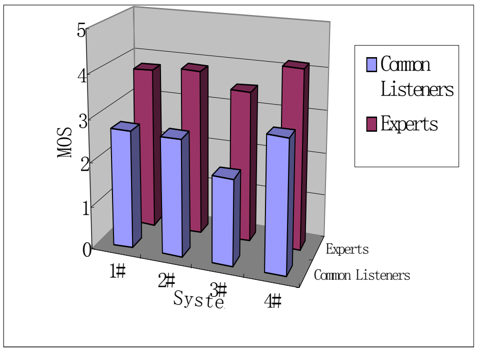
A diagram of the setup for assessment of speech synthesis systems

Table 2. Articulation scores of four text-to-speech systems

	System							
	1#		2#		3#		**4#**	
Score, %	Av	σ	Av	σ	Av	σ	**Av**	σ
S	76.4	5.4	80.4	4.6	75.0	6.2	*80. 1*	*4. 0*
W	67.1	7.7	73.0	7.6	75.8	9.0	*66. 9*	*5. 7*
J	86.0	10.4	82.4	9.5	83.3	14.2	*84. 3*	*10. 1*
JSUS	24.4		33.3		39.1		*23. 7*	

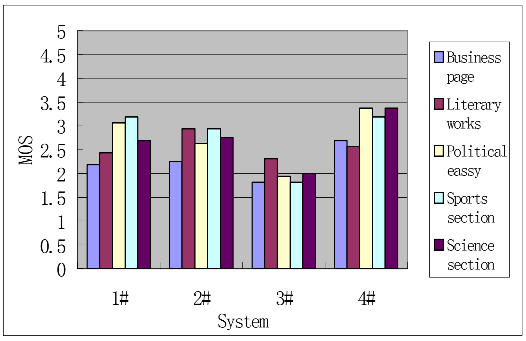
S: Syllable articulation, W: Word intelligibility, J: Sentence intelligibility, Jsus: SUS intelligibility.

Chart1. Speech naturalness in MOS for different styles of text of four text-to-speech systems



1#: syllable concatenation, 2#: PSOLA, 3#: PCM waveform, 4#: Cepstum parameters

Chart 2. Speech naturalness in MOS of four text-to-speech systems (Average of five styles of text)



1#: syllable concatenation, 2#: PSOLA, 3#: PCM waveform, 4#: Cepstum parameters

DISCUSSION

The speech synthesis assessment on network is proved very effective. We realized the automatization from the distribution of testing materials to speech waveform replay. It reduce the influence from some reason of artificial and unartificial, and let the results of the assessment be more accurate. Internet is very popular topic now , we can carry out the assessment on the internet. More synthesis systems can attend the assessment more easily. The assessment of 1997 will be a good preparation for the future.

REFERENCES

[1] Zhang, J., Qi, S. and Yu, G., "Assessment methods of speech synthesis systems for Chinese", Proc. ICPhS'95, Vol. 2, 206-209, 1995.

[2] Spiegel, M.F., Altom, M.J., and Macchi, M.J., "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech", Speech Communication, **9** , 279-291, 1990.

[3] Carlson, R., Granstrom, B. and Nord, L., "Evaluation and development of the KTH text-to-speech system on the segmental level", Speech Communication, **9** , 271-277, 1990.

[4] Zhang, J., "On the syllable structures of Chinese relating to speech recognition" Proc. ICSLP'96 Vol.4, 2450-2453, 1996.