TOTAL QUALITY EVALUATION OF SPEECH SYNTHESIS SYSTEMS

Jialu Zhang, Shiwei Dong, and Ge Yu*

Institute of Acoustics, Academia Sinica , *Institute of Software, Academia Sinica

Beijing 100080, CHINA

ABSTRACT

Based on the performance assessment of speech synthesis systems for Chinese the total quality evaluation of them has been carried out regular since 1994. The total quality evaluation includes speech intelligibility test at different levels (syllable, word and sentence), speech naturalness test and anti-interference ability test for phonetic module and text processing ability test for linguistic module. The designing principle of testing materials and testing methods are briefly described and the test results of four text-to-speech (TTS) systems for Chinese are presented in this paper. It is shown that 1. All professional technicians joining the speech naturalness test with the testing crew together overestimated the overall quality of the four tested systems; 2. The word intelligibility and Semantically Unpredicted Sentence (SUS) score are good for evaluating speech synthesis systems; 3. The anti-interference ability of synthetic speech is rather weak about 20 per cent in syllable intelligibility lower than natural speech under condition of S/N=5 dB.

1. INTRODUCTION

At present some national and international standards (Acoustics, GB/T 15508-1995) and recommendations related to the performance assessment of speech communication systems were established and a great improvement of speech communication systems was gained from these work. As a part of man-machine speech communication systems, speech synthesis systems and/or TTS systems are being developed along the similar way. There are, however, some fundamental differences between assessment of speech synthesis systems, especially TTS systems, and assessment of speech technologies, such as telephone systems and speech coding systems (Sproat, 1998).

The national assessment of speech synthesis systems for Chinese has been regular carried out in the mainland of China since 1994 (Zhang, 1995). This assessment activity mainly serves the developers of speech synthesis systems and aims at studying what is universal, if there is any, and what is language specific in speech quality evaluation of speech synthesis systems. Now there are different types of TTS systems being developed in the mainland of China although they have not come into the market yet.

This year, 1998, a new guidelines to assessment of speech synthesis systems, which aims at promoting the assessment to be standardizeable, automatizeable and web-based. Four different TTS systems were evaluated in March 1998. In this paper the testing setup is described in Section 2. Sections 3 and 4 devoted to issues in phonetic module test and linguistic module test and the speech naturalness test is explained in detail. In Section 5 the testing results are presented. Finally, we made some remarks for future research in Section 6.

2. TESTING SETUP

As a first step of building a web-based testing setup for evaluating speech synthesis systems, a testing center and a local network (10 BASE-T ether network) was established. All tested systems were connected with the testing center through the network. The testing materials including speech intelligibility test lists and texts for speech naturalness test and text processing ability test were generated for and sent to each tested system individually by using text generation tools. Then tested systems should make their response in two kinds of files---*Pinyin* (Chinese Transcription Alphabet) text files and synthetic speech wave form data files immediately and no manipulation by hands is allowed. The synthetic speech of each tested system was produced by the testing center with a sound adapter card of Sound Blaster 16. A certain pause for listeners make responses was inserted after a testing item according to the arrangement of test programs.

3. PHONETIC MODULE TEST

The principles we followed in designing the testing materials and methods are as follows:

(1) Representative—All intelligibility test lists are phonetically balanced and consistent with the phonotactic structures of the language. Each testing item (one item includes several testing elements) is embedded in a carrying phrase to present to the listeners. So that both the production and the perception sit in the real environment of continuous speech.

(2) Equivalent—Generally, the difference in intelligibility scores of two lists for one system is less than the standard deviation among listeners. But at sentence level it is difficult to satisfy this requirement.

(3) Sensitive—All testing lists are open set so that the intelligibility score of random choice is very low, and the segmental intelligibility hardly appears to be saturation under different conditions.

(4) Diagnosable—The diagnostic information of the tested systems can be obtained by using the perceptual confusion analysis of speech sounds.

(5) Efficient—The total testing time of a testing list is no longer than 5 min. (syllable list) or 15 min. (word list).

These intelligibility test lists were edited and re-balanced several times based on the experimental results under different conditions in past years. And some satisfactory results were obtained in the last two times of evaluations of speech synthesis systems (Zhang, 1995).

3.1. Testing crew

Sixteen young students (8 male and 8 female) with normal hearing, they are native of Beijing, organized into the testing crew. They have not experienced synthetic speech and were given some instruction and training for four hours before formal evaluations.

3.2. Syllable intelligibility test

Each phonetically balanced syllable list includes 75 syllables which are divided into 25 three-syllable groups (items) at random. And each item is embedded in a carrying phrase.

3.3. Word intelligibility test

The phonetically balanced word lists were produced by a word list generator. The words were selected from a labeled dictionary of which the vocabulary size is about 50,000 words. Each word list has 100 words (25 monosyllabic, 65 disyllabic and 10 polysyllabic, the distribution is in accordance with general text), they are divided into 25 four word groups to present to the listeners. Each four word group is a syntactically correct but semantically anomalous sentence, i.e. the Semantically Unpredicted Sentence (SUS), such as *Xigua chi heisede taiyang* (The watermelon eats the black sun). Before each group a series number is added when it is presented to the listeners. The statistical relation between the word intelligibility and the SUS score is investigated in Section 5.

3.4. Sentence intelligibility test

A sentence list has 25 simple sentences they were selected from newspapers. Generally there are no more than seven words in a sentence, in order to reduce the listeners' load. A series of experimental results show that the deviation of intelligibility scores is larger than syllable and word and simple sentences are not enough for evaluating TTS systems in general.

3.5. Speech naturalness test

Some important phonetic issues which can not be properly examined by intelligibility test, such as the r-colored finals, the words with neutral tone and some special tone modifications and tone sandhi rules, are evaluated in this stage. And we set the natural speech as reference so that all phonetic and linguistic and paralinguistic features are evaluated as a whole.

According to our experiences, speech naturalness measured in MOS is somewhat influenced by some linguistic and psychological factors, such as the text genre and presentation sequence of the output of TTS systems. This time both the testing material and the testing method were redesigned. The testing text was selected from different sections—1. political essay, 2. literary works, 3. sports sections, 4. science sections, 5. business pages of newspapers. Then the text selected was revised and some new contents, according to the requirements of text processing ability described in Sections 4.1 to 4.4, were added manually to make the text to be good for coverage and more linguistically dense, in order that the testing time is not so long.

The presentation sequence of the output of tested systems and sections of text was arranged by the testing center at random. The listeners were asked to evaluate the overall quality of the synthetic speech in a five point scale MOS and they gave a score after each section presented. One of two additional marks, + or -, can be attached to the score if it is necessary. The average score of the five sections for each tested system was considered as the final Mean Opinion Score.

It is worth to mention that in order to do comparative study of the influence of experience in synthetic speech, in addition to 16 common listeners four experts working on speech synthesis joined the evaluation of speech naturalness.

4. LINGUISTIC MODULE TEST

At present we mainly pay attention to the grapheme-to-phoneme conversion in TTS system for Chinese. It is well known that Chinese is a tone language with multi-tone system and Chinese characters are a kind of ideogram. It is important to examine the text processing ability of TTS systems for Chinese.

4.1. Word parser

All Chinese texts are in Chinese characters from historical literature to modern publications and newspapers. And no word boundary appears in the text. So we do need the word parser instead of the morpheme parser in European languages.

4.2. Transformation of number strings

Chinese has a unique numeral system. Generally in Chinese text numbers with higher than two places are written in Arabic numerals and they should be read in accordance with the Chinese numeral system. Especially some numbers (*yi* (one), *qi* (seven), and *ba* (eight)) have specific tone sandhi rules when they are followed by another word.

4.3. Homograph and polyphony character processing

It is well known that a Chinese character is a syllable and there are only about 1,200 different tonal syllables being used in spoken Chinese but more than 40,000 different characters in written Chinese. On the contrary some characters are polyphonies, their proper pronunciation can only be determined in context.

In addition, words with neutral tone syllable and some syllables with r-colored finals should be carefully processed at both word and sentence levels.

4.4 Symbols

The punctuation marks, metric units, e-mail address and some letters and symbols (for example, in foreign currency) in text should be pronounced in proper way.

The testing text was produced in the same way of text for speech naturalness test, there were two hundreds of testing points in a text of about 2,000 characters. The text generation tools are being developed.

5. RESULTS

5.1. Speech intelligibility test

syst.	1#		2#		3#		4#	
	Av	σ	Av	σ	Av	σ	Av	σ
S	74.6	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

Note: 1. Systems: 1#–syllable concatenation; 2#–PSOLA; 3#–PCM-coded; 4#–hybrid. 2. S–syllable intelligibility; W–word intelligibility; J–sentence intelligibility. Av=Average, σ=standard deviation.

Table 1. The results of intelligibility test of four tested systems.

From Table 1. it can be seen that the differences between syllable intelligibility of different systems are not so big for they are all based on syllable concatenation. According to our experimental results of natural speech under different transmission conditions the sentence intelligibility J will reach to 100 per cent when syllable intelligibility S is higher than 75 per cent. Now the sentence intelligibility of the four tested systems still paces up and down around 85 per cent , this is maybe due to the unsatisfactory processing of prosody and/or the perceptual multi-dimensionality of synthetic speech.

5.2. Speech naturalness test

syst.	1#		2#		3#		4#	
	Av	σ	Av	σ	Av	σ	Av	σ
MOS	2.71	0.75	2.70	0.82	2.0	0.68	3.0	0.71

Table 2. The speech naturalness in five point scale MOS of four TTS systems for Chinese. Av=Average, σ=standard deviation.

Comparing Table1. with Table2, it can be seen that there is no close correlation between speech intelligibility and naturalness. And it is worth to notice that the developers of speech synthesis systems always overestimate the overall quality (speech naturalness) of synthetic speech, see Fig. 1. Perhaps they attach some subjective weights to the overall quality for they have experienced hard time in improving the performance of speech synthesis systems.

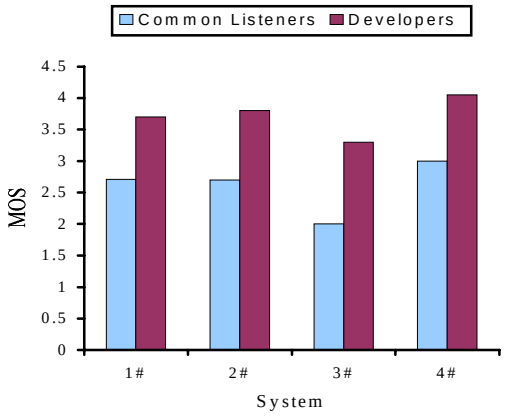

Fig. 1. The speech naturalness in MOS of four TTS systems evaluated by 16 common listeners and 4 developers.

5.3. The statistical relation between the SUS score and the word intelligibility

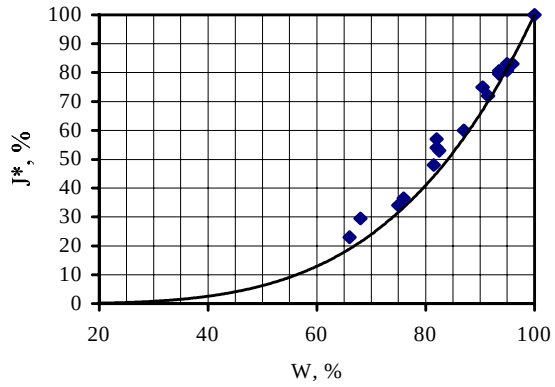

Fig. 2 The statistical relation between the SUS score J* and the word intelligibility W.

The word lists are presented in the form of semantically unpredicted sentences during intelligibility test. And two kinds of intelligibility score (word Intelligibility W and SUS score J*) were counted. The statistical relation between the SUS score and the word intelligibility for different speech synthesis systems (7 systems in 1994, 8 systems in 1995, 4 systems in1998) is drawn in Fig. 2.

If the words in the Semantically Unpredicted Sentences are perceived independently then we have the relation

$$J^*=W^4 \quad (1)$$

The theoretical relation–formulas (1) was drawn in solid line and the experimental results were drawn in black diamonds on Fig. 2. It can be seen that the experimental results fit in with the theoretical relation quite well.

5.4. The text processing ability test

The testing results of text processing ability for four TTS systems for Chinese are shown in Table 3. It can be seen that the error rates in text processing for all four tested systems are quite high especially for system 1# which has no r-colored final rules.

System	1#	2#	3#	4#
Err. Rate (%)	39.7	14.5	18.2	20.6

Table 3. The error rates of text processing for four TTS systems for Chinese.

5.5. The anti-interference ability test

The anti-interference ability of synthetic speech was tested under different signal-to-noise ratios. The sound pressure level of synthetic speech was fixed at 70 dB and the S/N was changed by changing the white noise level. The testing results were listed in Table 4. In order to make some comparative studies the testing results of natural speech under the same conditions were shown in Table 4, too.

System	1#	2#	3#	4#	Natural
S/N>30dB	76.4	80.4	75.0	80.1	98.7
=15dB	47.7	73.1	57.4	65.6	96
= 5dB	13.3	45.7	51.1	40.8	70

Table 4. The syllable intelligibility S of synthetic speech and natural speech under different signal-to-noise ratios(S/N).

From Table 4 it is obvious that the intelligibility of synthetic speech is degraded rapidly as the noise level increased and the anti-interference ability of synthetic speech is rather weaker than natural speech.

6. REMARKS

The fact segmental intelligibility is quite high for some speech synthesis or TTS systems is not enough for practical application in real world. Because the naturalness of synthetic speech is still not satisfactory and the anti-interference ability is rather weak. However, some developers are easy to be intoxicated with self-satisfaction of the overall quality of synthetic speech.

The multi-dimensionality of synthetic speech brings serious methodological problems to performance assessment research. Now synthetic speech is not at the same level of prosodic characteristics as natural speech. As Benoit said, the human is at the center of our investigation (Benoit, 1997). First of all we have to gain deep insight into the differences between the perceptual features of synthetic and natural speeches.

There is no doubt about it that improvement of prosodic characteristics will do great for performance of TTS systems especially for Chinese, perfect prosodic characteristics can help increase the segmental intelligibility. As Pisoni pointed out that "prosodic characteristics are not perceived directly by naive listeners; rather, they exert their influence indirectly on the processes used to recognize words and understand the meaning of sentences and discourse"(Pisoni, 1997).

A lot of work had been carried out on the Articulation Index (AI) (Zhang and Ma, 1965) and Rapid Speech Transmission Index (RASTI) for Chinese, but we have done nothing about the objective evaluation and prediction of intelligibility of synthetic speech.

7. ACKNOWLEDGMENTS

This research was supported by the National Hi-Tech Project 863 on the contract No. 863-306-03-09-02.

8. REFERENCES

1. ACOUSTICS-Speech articulation testing method, National Standard of China GB/T 15508-1995.

2. Benoit, Christian. “Evaluation inside or assessment outside?”, in *Progress in speech synthesis*, edited by van Santen, Jan et. al.,Springer, 1997.

3. Pisoni, David B. “Perception of synthetic speech”, in *Progress in speech synthesis*, edited by van Santen, Jan et al., Springer, 1997.

4. Sproat, Richard. *Multilingual text-to-speech synthesis :The Bell Labs approach*, Kluwer Academic Publishers, 1998

5. Zhang, J. and Ma, D. “A new method for deriving Articulation Index”, *Acta Acustica*, Vol.2, 80-84, 1965.

6. Zhang, J., Qi, S., and Yu, G. “Assessment methods of speech synthesis systems for Chinese”, *Proc. ICPhS’95*, Vol.2, 206-209, 1995