ASSESSMENT METHODS OF SPEECH SYNTHESIS SYSTEMS FOR CHINESE

Jialu Zhang, Shiqian Qi and Ge Yu
Institute of Acoustics, Academia Sinica
Beijing P.O.Box2712, China

ABSTRACT

A national assessment of the performance of speech synthesis systems for Chinese has been carried out yearly since 1994. The synthetic speech quality of five different systems were evaluated and diagnosed by using speech intelligibility tests. 16 college students (8 male, 8 female) with no experience with speech synthesis were the listeners, they were asked to do open response task by pencil-paper. In addition, speech naturalness was measured by Mean Opinion Score(MOS) on a ten Point scale. The perceptual confusion matrices of consonants were analyzed in order to give some diagnostic information of the synthesis systems at segmental level. And the statistical relations between speech intelligibilities at different levels of synthetic speech were compared with that of natural speech to explore some deficiencies in prosody. It is shown that stable and rational quality index and diagnostic information can be obtained in this method.

The evaluation methods of prosody have not been become available.

INTRODUCTION

With the growing use of synthetic speech in information-retrieval and man-machine communication systems in China, speech synthesis systems including text-to-speech systems are developed rapidly. Regarding to the COCOSDA work[1,2,3,4] in this field a national evaluation of the performance of synthesis systems for Chinese is to be needed in order to promote the development and the enhancement of speech synthesizers and to compare them with the systems for different languages. The Intelligent Computers Section of National Project 863 and the National Natural Science Foundation have supported the annual assessment activity on speech synthesis and recognition systems for Chinese since 1994.

So far more of available text-to-speech systems (TTS) have a phoneme intelligibility score close to that for natural speech and none of the present TTS for Chinese equipped with complete prosodic rule systems. At first stage we still have to pay more attention to speech intelligibility. Speech intelligibility test method for Chinese was developed at the beginning of 60s and it became one of the national standards.

The goal of the test is to evaluate the speech quality of different synthesis systems and to give some diagnostic information for each system individually.

Chinese is a tone language with multi-tone system and there are some distinctions at both phonetic level and syntactic level. For instance, there are only about 400 mono-tonal syllables used in real speech, almost no affixes exit at word level and the principle of sentence structures is almost the same with the phrase structures. So we have to carefully design the testing material according to the phonetic and linguistic characteristics of Chinese. The testing materials are described in section II and section III explains the testing methods. Then the results and discussion are presented in section IV and V.

TEST MATERIALS

The syllable lists

The syllable lists KXY1-10 are phonetically balanced in initials, finals and lexical tones. Each basic syllable list has 75 syllables divided into 25 trisyllablic groups. The basic syllable lists totaled 10 (KXY1-10). A test item is randomly composed of three syllables, so that a lot of testing syllable lists can be multiplies greatly from the basic lists. And there is no memory effect in the lists. Every test item---a trisyllabic group is added to a carrier sentence, such as "Dì yī zǔ shì ×××" (The first group is ×××.)to present to the listeners.

The word list

The word lists KXC1-10 are phonetically balanced too. There are 100 words in each word list in which the occurrence frequency of the word with different lengths is nearly the same with daily speech. There are 10 basic word lists composed of 1,000 common words, which were elected from "A list of 3,000 word of standard Chinese"[5] by 30 people.

Each list was divided into 25 four-word groups when it is presented to the listeners. A four-word group is a semantically unpredicted sentence, such as "Xīguā chī hēisède tàiyáng" (The watermelon eats the black sun).

By changing the four-word combinations a lot of different testing word lists can be deduced.

The sentence lists

Each sentence list is composed of 25 sentences which were selected from newspapers. Generally the sentence length is less than 7 words, and the content of the sentence lists covers a variety of domains. A strict criterion for judging the sentence intelligibility adapted is that any one key word i.e. the content word in the sentence is mistaken then this sentence is considered to be wrong. So that none of the poor systems can easily get 100% score with sentence lists

TEST METHODS

Listeners

16 college students (8 male, 8 female) who speak standard Chinese---"*putonghua*" with normal hearing were selected as the listeners. They have no experience with synthetic speech. Some instructions and training were given before test work. Especially, it should be explained that the tests aim only at to evaluate the synthesis systems not at the listeners, and the scores are only of the indices of the system performances. They were asked to do an open response task to write down the syllables, the words or the sentences on a special form after the stimuli presented.

Training

The listeners were trained by doing speech intelligibility tests with natural speech and with degraded speech (narrow bands and lower S/N), for 2-4 hours before evaluating the synthesis systems. And the test materials read by two speakers(one male and one female), which were included in the CAS speech database, were presented at first during test sessions, in order to set a reference of making naturalness assessment.

Testing

Two syllable lists, two word lists and two sentence lists were presented to the listeners sequentially, and the listeners were asked to make qualified judgments after each word list and each sentence list on a 10 point scale of naturalness. The categorical judgment on the 10 point scale is like that : excellent -9~10, good-7~8, fair-5~6, bad-3~4, very bad-1~2.

It took about 35 min. to evaluate a system and then a 5 min. break was given for listener rest and tested system preparation.

The listeners wrote down their open response in Chinese characters or in "*Pinyin*"(Chinese phonetic alphabet).

RESULTS

The speech intelligibility and naturalness of five systems with normal speech rates (3~4 syllables/s) are given in Table 1 and 2.

1. Speech intelligibility

Two lines of data of intelligibility scores S, W and J for each system were resulted from using two different test lists respectively, $\bar{x}$ stands for the average value over all listens and σ the standard deviation. In Table 1 the fact that the differences in intelligibility S and W between two test lists in the same category are less than the deviation among listeners means that the test lists are well equivalent. As for the sentence intelligibility J the deviations are rather high.

It is worth to notice that the sentence intelligibility J of natural speech generally is higher than the word intelligibility W. For synthetic speech of the systems tested the

relation is inverse except system 4#. That means on the one side all the systems evaluated are not so good in prosodic processing and the judgment criterions of sentence intelligibility we adopted are sensitive and practical on the other.

Table 1, Intelligibility scores of speech synthesis systems for Chinese

Interlligibility		Natural Speech		1#		2#		3#		4#		7#	
		$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Consonant	C			86.7		80.0		80.8		72.5		96.5	
Syllable	S	95.6	3.1	80.7	6.5	73.6	4.3	74.5	6.0	65.6	6.5	94.9	2.3
		94.0	3.7	83.9	5.7	76.0	3.8	77.6	5.6	66.2	9.1	93.4	2.8
Word	W	98.1	1.6	94.3	3.2	87.0	4.6	76.3	8.0	81.8	6.5	95.1	1.5
		97.6	2.4	92.9	5.1	79.8	5.0	73.4	8.1	81.8	7.2	96.3	1.8
Sentence	J	99.9	0.7	57.0	13.5	54.1	14.6	61.5	13.8	83.3	10.1	89.5	7.4
		100	0	74.5	10.7	75.5	12.7	60.0	7.7	86.1	9.0	95.0	5.4

Note: 1# - PCM waveform edited; 2# - CELP concatenative; 3# - VQ-LPC concatenative; 4# - formant; 7# - PSOLA.

2. Naturalness

The naturalness in MOS of the synthetic speech of five different synthesis systems is listed in Table 2.

Table 2. The naturalness in MOS of five synthesis systems for Chinese .

	Systems				
	1#	2#	3#	4#	7#
Naturalness	4.6	3.2	3.4	5.6	7.8

It can be seen from Table 1 and 2 that the naturalness of synthetic speech is still not so high although some systems, such as 7#, can get quite high intelligibility. The ,speech quality of naturalness 7.8 in MOS, the highest in the five systems, is nearly equivalent to that of 8k bps VSELP (Victor Sum Excitation Linear Prediction) telephone system.

3. Comparison between natural speech and synthetic speech

Fig.1 shows the statistical relation between syllable intelligibility and word intelligibility and Fig. 2 the statistical relation between syllable and sentence intelligibities, they were established for natural speech transmitted under different conditions. And the data of synthetic speech of five systems were also drawn on Fig.1 and Fig.2.

From Fig.1 it can be seen that the statistical relation between syllable intelligibility and word intelligibility is almost the same for natural speech and synthetic speech. As for

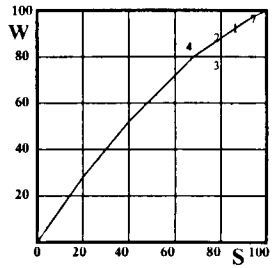

Fig.1 The statistical relation between syllable intelligibility S and word intelligibility W. (Solid line: natual speech; Numeric: synthetic speech.)

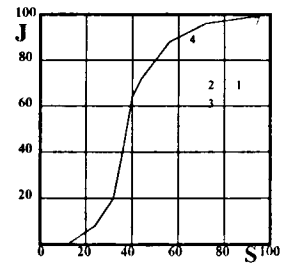

Fig.2 The statistical relation between the syllable intelligibility S and sentence intelligibility J. (Solid line: natural speech; Numeric: synthetic speech.)

the statistical relation between syllable intelligibility and sentence intelligibility in Fig.2, however, the sentence intelligibility of synthetic speech is always lower than that of natural speech. System 1# a text-to-speech system with bad prosodic rule system got lower sentence intelligibility and lower naturalness but higher syllable and word intelligibilities. On the contrary system4# a text-to-speech system using formant synthesizer with better prosodic rule system got higher word and sentence intelligibilities but lower syllable intelligibility.

4. The perceptual confusion of consonants

The perceptual confusion metrices of consonants were analyzed to get the diagnostic information.

In systems 1# and 7# the most frequent confusion was occurred in between voiceless fricatives and aspirated plosives and affricates. This phenomenon makes clear that the waveform edited systems, both PCM(1#) and PSOLA (7#), have some defects in voiceless fricative processing. As for the formant synthesizer, system 4#, the unaspirated plosives and affricates got the lowest segmental intelligibility, that means it is difficult to control the parameters of voiceless plosives and affricates by using formant synthesizers.

DISCUSSIONS

Speech intelligibility tests used as a assessment method of speech synthesis systems can give very much useful information about the system performance at phonetic level. It is more important at the developing stage in the laboratory. The perceptual confusion matrices of consonants give you a clear pattern of the defects the synthesis systems have and it can be a useful diagnostic method.

Almost all synthesis systems developed in China are syllable-based as there are only about 1200 syllables with tonal patterns being used in real speech. The tone intelligibility is high, and the major perceptual confusion of tones was in between rising tone and dipping tone. The same confusion pattern of tone perception was observed in natural speech[6] , too. Maybe this is the psychological basis of the distinct tone sandhi rule --- the dipping tone should be changed into rising tone when it is followed another dipping tone in spoken Chinese.

There is another lexical tone --- light tone at word level in spoken Chinese besides the four lexical tones at syllable level. The light tone syllable is really atonic from the acoustical point of view, but it was treated as a toneme in traditional phonology of Chinese. From Fig 1 it can be seen that the word intelligibility is closely related to the syllable intelligibility, and the word intelligibility is not so strongly effected by the prosodic features. At sentence level, however, thing are different.

REMARK

There is no doubt that prosodic features, especially the interaction between tones and intonation, are more important for speech naturalness of Chinese but the evaluation methods have not become available. A lot of work have to be done in this field.

REFERENCES

[1] Pols, L.C.W. and SAM-partners, "Multi-lingual synthesis evaluation methods", ICSLP'92(Banff), Vol.1, 181-184, 1992.

[2] "Section Two: Synthesis Assessment" in Proc. 1992 Workshop of COCOSDA (Banff), Ed. by Kate Jones and Joseph Mariani, PP.II:1-II:4, 1992.

[3] "Synthesis"in Report on the COCOSDA Workshop (Berlin), PP.25-28, 1993.

[4] "Synthesis" in 1994 International workshop on International Coordination of Speech Databases and Speech I/O Systems Assessment, PP.15-21, 1994(Yokohama).

[5] "A list of 3,000 words of Standard Chinese" Ed. by Research and Spread Section, China Charater Reform Committee, Character Reform Publishing House, 1959 (Beijing).

[6] Yang, Y. and Fang, Z., "Study of Tone Perception of Standard Chinese", in Proc. 5th International Symposium on Chinese Language - Cognition Science, (Science Publishing House, Beijing), 1992.