ABSTRACT
This study tries to put forward a kind of highly effective evaluation method according to the actuality of synthetic Mandarin. Through evaluating the two Mandarin TTS system of Beijing info-quick Sino-voice speech technique corp., the paper explores MOS (Mean Opinion score) method with a view of practicality. During operating, we defined the five-score rather accurately, and composed a set of testing materials. In addition, we trained the listeners on the basic knowledge of phonetics. The listeners participated in the judgment of speech of each score. Therefore, it increased the consistent understanding among different listeners, and avoided excessively disperse scores of different listeners. Moreover, during testing, the materials of two TTS system were set at random, and it can avoid learning effect for listeners. Besides MOS estimate, we also tried to use rather objective evaluating method. The result of assessment is reasonable.

1. INTRODUCTION
There are some mature studies for the evaluation method of synthesized speech. Mainly categorized as: global testing, diagnostic testing, objective testing and application-orientated testing. The frequently used method in global testing is Mean Opinion Score (MOS)[1] Objective evaluation method for Mandarin speech synthesis system are studied in recent years[2] We mainly used MOS evaluation method in this paper.

With the development of Mandarin speech synthesis technique, the exploring for evaluation method catches sight of researchers. In recent years, 863 workgroup of speech synthesis evaluation obtained some achievements. At the same time, there is a lot of room to improve. For the same system, the results of evaluation in different time are varied largely. There is no uniform standard and shortage of phonetics knowledge training in 863-synthetic-speech-valuation. Because multi systems are evaluated successively in the same time using the same testing materials, there are learning effect existing in the later evaluation. And because the listeners are short of phonetics knowledge training, the evaluation results tend to be subjective and random. In this study, we lay particular stress on the followings. First, we put forward the evaluation method emphasized on intelligibility and naturalness according to the actual synthesized speech using concatenation technique. Second, we pursue practicality on defining the five-score and choosing the standard speech of each score in order to achieve unity of different evaluations. Third, the testing materials of different systems were reset at random to preclude the listeners from learning effects. Fourth, listeners were trained on the related phonetics knowledge and speech synthesis technique to lessen the spread of scores among listeners.
2.1 The Definition of Five-score
Although the five-score belongs to a kind of subjective evaluation, it is necessary to be maneuverable and sensitive to the quality of present synthetic speech.

Frequently used technique of actual speech synthetic speech system is concatenation. On the articulation of syllable it has already no problem, the focus is on the naturalness that is essential for us to distinguish different synthetic speech systems. We adopted frequently used five-score evaluation method (MOS) on evaluation of sentence and paragraph naturalness. Consequently, we should pay attention to accurate and maneuverable definition of five-score.

The following is the specific definition.

Score 5 (excellent): the ideal score 5 is broadcasting speech. The score 5 of synthetic speech is similar to broadcasting speech. It is difficult to distinguish them. On the whole, it sounds very well, clear, fluent*. Listeners take pleasure in it.

Score 4.5 (to be popularized): clear, understandable, fluent, sounding well on the whole, listeners are willing to accept it (they will not feel tired within half an hour). No obvious prosodic errors. It can be popularized.

Score 4 (good): Although listeners make do with it, they can understand it without problem. Without obvious word-dividing errors and serious faults in rhythm, text-to-speech** is generally right. There are some shortcomings in stress, and one or two syllables are not clear. Listeners rarely endure it over thirty minutes continuously.

Score 3 (fair): It is generally understood, but text-to-speech has obvious shortcomings, and there are many faults in rhythm and influence between syllables, so listeners are unwilling to accept it. Due to the obvious tiredness, it is difficult for listeners to stand up with ten minutes continuously.

Score 2 (bad): Listeners cannot hear clear some keywords. The kind of speech is similar to concatenation of monosyllables crudely.

Score 1 (very bad): the speech that is machine-quality sound is not clear, and cannot be understood. It can only express broken or little information. It is difficult for listeners to guess meaning. It cannot be accepted at all.

* “fluent” refer to reaching the level of pleasing to the ear in prosody and co-articulation.

** polyphone-characters, neutral and retroflex, string of number, the pronunciation of English words.

2.2 The Training to Listeners
During the evaluation process of synthetic speech of info-quick Sino-voice corp., there are ten listeners participating in testing totally. The training to listeners is another important item. While planning evaluation program, we attained common agreement, which is that we should train the listeners on necessary phonetics and speech synthesis technique knowledge in order to gain consistent and objective evaluation.

Before formal testing, there are two trainings and simulation judgment of naturalness for the listeners.

The two trainings are made up of phonetics knowledge and speech synthesis technique. The phonetics knowledge training includes the concept of Chinese syllable formation, intonation and stress. Consequently, the listeners can not only understand speech in view of sensibility, but also evaluate speech rationally. The training of speech synthesis technique made listeners know the history of speech synthesis: from mechanism synthesis, formant synthesis to present concatenation driven by large speech database. The listeners realized that the synthetic speech is not equal to nature speech and has its own characters.

Listeners’ simulation judgment of naturalness is our focus during the training. We offered 100 sentences including synthetic speech that is synthesized by different methods and different fluency natural speech.
We demanded the listeners evaluate the 100 sentences according to the definition of five-score. We worked out results of the evaluation statistically. For each sentence, there are average number and standard deviation of scores. We chose the sentences whose average number is close to 1 and whose standard deviation is very little as the standard speech of score 1. According to the above principle, we gained the speech samples of score 2, 3, 4, 4.5, 5. Then, we made listeners familiarize all the speeches repeatedly. And made them realize the standard of 5-score in view of sensibility and relatively consistent understanding.

2.3 Testing Contents and Design of Testing Materials
Besides naturalness, the intelligibility of sentence is also important standard of evaluating synthetic speech. The naturalness is the key point for synthetic speech whether to be popularized, and the intelligibility is the key point for synthetic speech whether to be used. Either of them cannot be missing. Therefore, our testing contents mainly include the naturalness of sentence and paragraph and the intelligibility of sentence. In addition, the articulation of syllable is also taken as an item of testing.

The testing of intelligibility adopted the method of dictation. The sentences dictated by listeners are permitted of errors of lacking or adding some characters by mistakes, but these errors cannot affect the right understanding of sentences.

The testing materials of syllable articulation are made up of two tables specially devised for syllable articulation testing, which are chosen from GB/T 15508[3] 1995 Acoustics-Speech articulation testing method[3]. The testing materials of intelligibility and naturalness are chosen from media texts, mainly from people daily. In addition, there are some special materials for prosodic division and traditional testing material North Wind and Sun. The length of each sentence for intelligibility testing is within 15 syllables. The intelligibility testing materials for one system are 20 sentences. The sentences of the two systems are not same, but the difficulty and testing points are similar. The length of each sentence for naturalness testing is within 25 syllables, and the sentences used by two systems are same. The length of each paragraph is within 200 syllables. Both of the two systems use the same 3 paragraphs.

As to designing testing material, we try to devise testing points into the materials, which cover with essential speech phenomenon. The accuracy ratio is worked out finally. The testing points are generally divided into three groups. The first one is the pronunciation of number, English and polyphone characters. The second one is neutral tone, retroflex, co-articulation and tonal modification. The third is accuracy ratio of word-dividing or syntax processing. For example, the sentence of $\text{国民经济增长速度高达10\%左右}$ includes the pronunciation of number 10% and polyphone character 长. The sentence of $\text{APEC 上海将在这儿举行高层聚餐}$ includes the pronunciation of English word APEC. The sentence of $\text{土改运动中他显示了自己的本事}$ includes tonal modification of 土改 and retroflex 本事. The sentence of $\text{他只对比他强的选手有兴趣}$ includes tonal modification of 对 and neutral word 比. Whether the two words 对 and 比 are wrongly processed as a word 比对 in the sentence of $\text{他只对比他强的选手有兴趣}$, which belongs to the part of word-dividing or syntax processing. The testing points method is close to objective evaluation. The examiner judges the testing points whether to be right according to testing points’ note. Then, the accuracy ratio of testing points can be worked out, which is an important part in testing results.

In the parts of intelligibility and naturalness, testing materials synthesized by two different Chinese TTS (Text-To-Speech) were reset at random in order to preclude the listeners from learning effects. While computing the results of evaluation, the scores of two systems should be separated.

3. RESULTS AND DISCUSSION
3.1 The Results of Evaluation

The system driven by large speech database (size: 1.92G):
- Articulation of syllable: 88.17%
- Intelligibility of sentence: 99.00%
- Naturalness of sentence (5-score): 4.34
- Naturalness of paragraph (5-score): 4.35
- Accuracy ratio of testing points: 89.16%

The system driven by small speech database (size: 1.0M):
- Articulation of syllable: 74.00%
- Intelligibility of sentence: 97.83%
- Naturalness of sentence (5-score): 4.09
- Naturalness of paragraph (5-score): 4.26
- Accuracy ratio of testing points: 86.59%

3.2 The Importance of Score 4.5 in Five-score

MOS (Mean Opinion Score) generally adopted five score scales, which does not specialize some scale with decimal. But, during evaluation, we found that people are sensitive between 4 and 5, and that score 4.5 is a standard of acceptability. Therefore, we claim that it is very important for score 4.5 to be threshold value of being popularized while evaluating synthetic speech.

3.3 The Objectivity of Testing Points

In this study, we adopted the accuracy ratio of testing points as an important standard of evaluating synthetic speech. Testing points generally cover with essential speech phenomenon in Mandarin. The judgment of testing points’ accuracy ratio doesn’t need subjective evaluation on the whole. For instance the word ""，if speech synthesis system processes "" and "" as one syllable of retroflex, the examiner will consider that the realization of this testing point is right. The judgments of different examiners should be consistent. In view of this, we can consider that the accuracy ratio of testing points is more close to objective judgment that is not similar to MOS.

4. CONCLUSIONS

This paper explored a kind of testing method for Mandarin TTS system that mainly adopted MOS evaluation method. This kind of method focuses on maneuverability and practicability. The following is features of this testing method: putting forward rather specific definition of five score; claiming that listeners should be trained in related phonetics knowledge and speech synthesis technique, and that the devising of testing materials should cover with essential speech phenomenon in Mandarin, such as neutral, retroflex and tonal modification.

This study also tried to use judgment method that is close to objective evaluation. The combination of objective and subjective evaluation will be our next work.

REFERENCES