



Sentence Level Intelligibility Evaluation for Mandarin Text-to-Speech Systems Using Semantically Unpredictable Sentences

Jian Li¹, Dmitry Sityaev², Jie Hao¹

¹ Research and Development Center, Toshiba (China) Co., LTD.

² Cambridge Research Laboratory, Toshiba Research Europe Limited

{lijian, haojie}@rdc.toshiba.com.cn, dmitry.sityaev@crl.toshiba.co.uk

Abstract

Intelligibility assessment is one of the important aspects in the text-to-speech system (TTS) evaluation. Several intelligibility assessment methods have been proposed and successfully applied to European languages, both at word level and sentence level. Since Mandarin has its own unique features, these methods must be modified when applying to Mandarin. The word level assessment methods such as DRT and MRT have successfully been modified and extended to Mandarin (e.g. CDRT, CDRT-tone and CMRT). Sentence level assessment methods, on the other hand, have not been well studied for Mandarin. This paper focuses on the Semantically Unpredictable Sentences (SUS) test, which is one of the most commonly used sentence level assessment methods, and considers several important aspects of the SUS test design when extending it to Mandarin. It also compares the SUS test for Mandarin with CDRT, CDRT-tone and CMRT.

Index Terms: TTS evaluation, intelligibility assessment, sentence level intelligibility, SUS

1. Introduction

Great improvements have been achieved in text-to-speech in recent years. Accordingly, various evaluations methods have been proposed to assess the quality of TTS systems. Assessing intelligibility is an important aspect in TTS evaluation. It tests whether synthetic speech can be correctly recognized by listeners.

The intelligibility of TTS systems is commonly evaluated at word level and at sentence level. DRT (Diagnostic Rhythm Test) and MRT (Modified Rhythm Test) have been widely used in TTS evaluation at word level [1], while the SUS (Semantically Unpredictable Sentences) test is the commonly used method for TTS evaluation at sentence level [2][3][4][5]. These methods have been successfully applied to European languages.

Mandarin is a tonal syllabic language and has its unique features. For this reason, it is not always straightforward to use the developed methods for European languages without any modifications. Some studies have been carried out to extend the word level assessment methods for Mandarin [1][6]. CDRT (Chinese DRT) and CMRT (Chinese MRT) were developed on the basis of DRT and MRT and take into account some features of Mandarin phonetics, whilst the CDRT-tone has been proposed to assess purely the intelligibility of tone itself. Although these methods have been applied in speech coding area [1][6][7], it is obvious that they can be used in speech synthesis area.

With word level assessment methods, isolated words are tested. For most TTS applications, continuous speech rather than isolated words are often synthesized. Since the prosody of continuous speech (i.e. duration, F0 and so on) is different

from that of an isolated word, it is necessary to research into the sentence level assessment methods for Mandarin. This paper focuses on the SUS test and its extension to Mandarin.

The rest of the paper is organized as follows. Section 2 gives a brief introduction of the SUS test. Section 3 introduces some unique features of Mandarin and brings up important points to be considered when designing the sentence level intelligibility test. In Section 4, the results of the experiments are presented. Section 5 provides the discussion of the results of the Mandarin SUS test and Section 6 offers some conclusions.

2. Brief introduction of SUS test

The SUS test is effective and reliable for intelligibility assessment at sentence level and it has been adopted by the SAM group for European languages [3], Blizzard Challenge 2005 [5] and 2006 [4] for English. This paper briefly introduces the recommendation of SUS test for European languages. Please refer to [3] for detailed information.

2.1. Basic ideas of SUS

As pointed out in [3], meaningful sentences contain a lot of semantic and syntactic contextual cues that often help understand some otherwise unintelligible parts of synthesized speech. Since the effects of these cues are not easy to measure, the meaningful sentences are not best for the assessment of a TTS system. In the SUS test, sentences are composed of words respecting the syntactic structure of a language. These words however do not have any relations to each other, so the produced sentences are semantically anomalous. Subjects can only get context cues of syntactic category but they cannot get any further information about the word identity. Thus the effect of semantic contextual cues is avoided.

The words composing the sentences are randomly selected from the lexicon. This removes the limit on the number of stimuli that can be generated. Potentially, a great number of sentences can be generated, which reduces the learning effect.

It is more difficult for subjects to remember these meaningless sentences than meaningful sentences. To ease the task, the sentences must obey the syntactic rules and the number of words in a sentence should be less than eight.

2.2. Syntactic structures

There are five simple syntactic structures used in the SUS test. These structures can typically be found in many European languages. These structures are simple enough that even non-native subjects can easily understand them.

The structures can be labeled by the syntactic categories. The words of each category are all frequently used words. Only if there are not enough monosyllabic words, should

polysyllabic words be used. The selection of the words in each category and some rules to prepare the words in the lexicon, such as the transitivity of the verbs, the tense of the verbs, etc, are also recommended in [3]

3. Features of Mandarin

We briefly introduce some unique features of Mandarin in this section. Important considerations when designing the Mandarin SUS test are also discussed in this section.

3.1. Chinese characters and Mandarin syllables

A Chinese character is a kind of ideographic character and a Chinese character is called *hanzi*.

In Mandarin, every *hanzi* constitutes a syllable. Each syllable has three integral parts to it: the initial, the final and the tone. Sometimes a syllable does not have the initial part and in this case, the initial part is called “empty initial”. Two syllables will be different even if only one part is different.

The Chinese Phonetic Alphabet Scheme [8] defines the alphabet for the initials and finals. The rules of combining the initial letters and final letters to form a legal syllable are also defined in the scheme.

In Mandarin, there are four lexical tones plus a special tone called “neutral tone”. In order to simplify the input of tones into a PC, digits from 1 to 4 are used to express the four lexical tones. For neutral tone, some researchers and systems use digit 0 while some others use digit 5.

The legal combination of initial letters and final letters followed by a tone is called *pinyin*. The *pinyin* can exhaustively define the pronunciation of a syllable. However, only educated people in China are able to read and write using *pinyin*.

3.2. Tone sandhi

In continuous speech, the tone of some *hanzis* may change. This phenomenon is often known as *tone sandhi*. For example, if the tones of the two *hanzis* in a word are both Tone 3, then the first Tone 3 is changed to Tone 2.

3.3. Relationship between *hanzi* and *pinyin*

In Mandarin, there are more than 1,300 *pinyins* [8] and a total of about 20,000 *hanzis*, out of which more than 6000 *hanzis* are frequently used. For this reason, in Mandarin several *hanzis* share the same *pinyin*. So if an isolated syllable is uttered, listeners cannot predict which *hanzi* is uttered.

Some *hanzis* have more than one pronunciation. They are called as heteronyms. The pronunciation of a heteronym in the sentence can only be determined according to the contextual information.

3.4. Chinese words

Each *hanzi* has a meaning and there are lots of *hanzis* in Mandarin that correspond to a one-*hanzi* word when used in the sentence. However, since most of the *hanzis* have more than one meaning, the actual meaning of a *hanzi* when used in the sentence can only be determined with the help of the contextual information.

In a vast majority of cases, a word is composed of two or more *hanzis* combined together. These words always have determined meanings and the possibility of all *hanzis* in a word of this type sharing the same *pinyins* with another word is very low.

Another feature of Mandarin is the use of auxiliary words. These words are attached to content words and are used to combine neighboring words into a phrase (respecting some syntactic structure), or to express the tense of the verb or certain grammatical mood. Although these auxiliary words do not have any actual meaning, they are still very important in Mandarin.

3.5. Important points to consider when designing the SUS test for Mandarin

3.5.1. Method to transcribe speech: *hanzi* or *pinyin*

There are a few disadvantages to be considered if subjects transcribe sentences using *hanzi*. Firstly, it is more complicated to write *hanzis* than to write using *pinyin*. Subjects will spend more time writing the same sentence using *hanzis* than using *pinyin* that also means subjects are very likely to forget the content quicker. Secondly, one pronunciation corresponds to several *hanzis*. Subjects may spend time to think which *hanzi* should be used for the pronunciation they heard. This means subjects may be distracted from concentrating on just listening to the speech.

To avoid the shortcomings mentioned above, it was decided to transcribe sentences using *pinyin*.

3.5.2. *Hanzi* number in word

Since one *hanzi* corresponds to one syllable and there are some one-*hanzi* words in Mandarin, it is possible to make up sentences entirely of one-*hanzi* words.

In the SUS test, all the sentences are meaningless, i.e. neighboring words cannot serve as semantic cues for the meaning of the word in question. Every word is thus recognized mainly by its pronunciation. In European languages, homophones are not so popular as they are in Mandarin. If a word is recognized, subjects can easily link the pronunciation with the concept that the word stands for and this link helps subjects to remember the word. In Mandarin, several *hanzis* can share a pronunciation, so it is therefore hard to link pronunciations taken out of context with any concept. For the SUS test in Mandarin, this will mean the subjects will only remember pronunciations. It also means that one-*hanzi* word task in Mandarin is a little more difficult than monosyllable word task in European languages.

Since most two-*hanzi* words always have the one-to-one relationship with their pronunciations, we hypothesize that subjects should normally be able to link the correctly recognized pronunciation (e.g. two syllables) with some concept. However, words consisting of two syllables provide a contextual cue for each which raises the question of the reliability of the results. It seems that one must consider having a trade-off between the difficulty and the reliability of the test.

3.5.3. Syntactic structures

The use of sentences with correct syntactic structures can help subjects to remember the sentence and thus ease the task. For Mandarin, it appears possible to follow the SUS test design and find some syntactic structures similar to those proposed for the European languages. We propose to consider two cases.

In Case 1, one-*hanzi* words are to be used. In this case, it will be harder for subjects to link words with any concepts.

Syntactic structures are not expected to contribute to the goal of easing the task.

In Case 2, two-hanzi words are to be used. As discussed above, pronunciations are more likely to be linked with concepts. In this case, it is expected that syntactic will contribute to the task of recognition. However, if every word in each syntactic category is a two-hanzi word, the total number of hanzis in each sentence increases which may cause an overload for subjects' short-term memory.

3.5.4. Auxiliary words

Auxiliary words also serve as important cues in decoding the syntactic structure. Using auxiliary words can therefore be beneficial (Cf. the use of articles in English), however, it means an increase in the total number of hanzis in each sentence that may affect subjects' short-term memory.

4. Experimental setup and results

We designed two experiments with a view to investigate the points raised in Sub-section 3.5. The design of stimuli for each experiment is described below. A Mandarin version of the TTS system described in [9] was used to synthesize the stimuli.

Eight subjects who were native speakers of Chinese took part in the experiments. There were 30 sentences in each experiment; the first sentence was used as a warm-up and not scored in the results. Each sentence was played only once to avoid the learning effect. After each sentence was played, subjects were asked to transcribe it using pinyin. All subjects took a long break between the two experiments.

Subjects were also asked four questions at the end of the two experiments: (1) Could you decipher the syntactic structure of the sentences in Experiment 1 and Experiment 2? (2) Does the syntactic structure help to remember the sentence? (3) Do auxiliary words help in deciphering the syntactic structure and thus making it easier to remember the sentence? (4) Is Experiment 1 easier than Experiment 2?

The transcribed results were automatically scored using HTK [10]. We scored the results separately: (a) by tone only and (b) by pinyin without tone. A sentence is considered to be transcribed correctly only if all the tones or pinyins without tone of all the hanzis in the sentence are transcribed correctly.

4.1. Experiment 1

In Experiment 1, a very simple syntactic structure was used: Subject - verb - direct object. Two types of sentences were generated for this structure: (1) noun + verb + adjective + noun and (2) noun + adverb + verb + noun.

In this experiment, two-hanzi words were used for every syntactic category. Auxiliary words were sometimes added after adverbs and adjectives. Some sentences did not contain any auxiliary words. The hanzi number in any one sentence was at least 8. Some sentences contained 9 or 10 hanzis. The results of Experiment 1 are presented in Table 1.

Table 1: *The result of Experiment 1*

Result		Corr	Sub	Del	Ins
Tone	Sent	67.08%			
	Hanzi	95.79%	3.65%	0.56%	0.28%
Pinyin (no tone)	Sent	69.05%			
	Hanzi	93.01%	6.43%	0.56%	0.28%

4.2. Experiment 2

In Experiment 2, three different syntactic structures were used: (1) noun + verb + preposition + adjective + noun, (2) adjective + noun + adverb + verb and (3) verb + noun + conjunction + noun.

In this experiment, only one word in each sentence was a two-hanzi word; all others words were one-hanzi words. Auxiliary words were also sometimes added after adjectives and adverbs in some sentences. The maximum number of hanzis in the sentence is 7 and the minimum number is 5. The results are presented below in Table 2.

Table 2: *The result of Experiment 2*

Result		Corr	Sub	Del	Ins
Tone	Sent	68.97%			
	Hanzi	93.60%	6.32%	0.07%	0.07%
Pinyin (no tone)	Sent	42.24%			
	Hanzi	85.17%	14.75%	0.07%	0.07%

5. Discussion

5.1. Extending the SUS test to Mandarin

Both experiments appear to achieve a high accuracy rate for recognizing pinyin without tone at hanzi level (93.01% in Experiment 1 and 85.17% in Experiment 2). In Experiment 1, all subjects said they could decipher the syntactic structures and almost all of them said that the structure did help them to remember the sentence. However, in Experiment 2, the subjects said they couldn't decipher the syntactic structure. This finding supports our hypothesis made in Sub-section 3.5.3, namely that syntactic structure can contribute to understanding. It is probably for this reason that all the subjects found Experiment 1 easier than Experiment 2, even though the sentences in Experiment 1 were significantly longer than the sentences in Experiment 2.

The results also reveal that in both experiments, some hanzis were inserted or deleted. The analysis of errors showed that in Experiment 1 most deletions occurred in long sentences made up of 10 hanzis. In Experiment 2, one insertion and one deletion happened in the sentence made up of 7 hanzis. This indicates that the number of hanzis used in Experiment 1 is probably large for subjects' short-term memory. Additionally, some subjects commented that 7 hanzis is about the maximum number of hanzis that they could remember. This number of hanzis was often used in Experiment 2.

With respect to auxiliary words in Experiment 1, all subjects but one said the presence of auxiliary words did not help them to decipher the syntactic structure or remember the sentence better. However, this one subject also mentioned that the auxiliary words were easily forgotten and she sometimes was confused whether she had actually heard the auxiliary word in a sentence or not.

Finally, the recognition accuracy for tone only in Experiment 1 (67.08% at sentence level and 95.79% at hanzi level) is very close to the recognition accuracy for tone only in Experiment 2 (68.97% and 93.60%, respectively). This demonstrates that the recognition of tone does not appear to be affected much by the difficulty level of the task.

5.2. The effect of subjects on result

In China there exist many dialects of Chinese. Mandarin is based on the dialect of the Beijing area and some parts of the North China. We will henceforth refer to this area as the Mandarin area and all other parts of China as the dialect areas. Educated people from the dialect areas can understand and speak Mandarin, however some of them are unable to distinguish certain phonemes. These include, for example, sibilated 'ch' and unsibilated 'c', nasal final 'in' and back nasal final 'ing', nasal initial 'n' and lateral initial 'l'.

We analyzed the errors for these phonemes by subject. Among the 8 subjects who took part in the experiments, 2 subjects were from the Mandarin area 6 subjects were from the dialect areas. In Experiment 1, the subjects from the dialect areas made a total of 43 errors, while the subjects from the Mandarin area made only a total of 3 errors for these phonemes. Similarly, in Experiment 2, the subjects from the dialect areas made 65 errors in total, and the subjects from the Mandarin area made only 3 errors. We further noticed that one subject from the dialect area contributed to 27 errors in Experiment 1 and 24 errors in Experiment 2.

These results reveal that certain phonemes of Mandarin present a challenge for the people from the dialect areas, which will ultimately affect their performance in a test like the SUS test. This should be borne in mind when selecting the subjects for the experiment and/or analyzing the result.

5.3. Comparison of the SUS test with CDRT, CMRT and CDRT-tone

CDRT, CMRT and CDRT-tone are methods for assessing intelligibility of Mandarin speech output at word level. In these three tests, syllables are organized into pairs or groups. The syllables in each pair or group are similar in pronunciation and differ only in one part of the pinyin. The subjects hear one syllable in the pair or group played to them and their task is to determine which syllable in the pair/group is played to them.

The results of the experiments we conducted show that similarly to the CDRT, CMRT and CDRT-tone, most of the incorrectly transcribed syllables in the Mandarin SUS test are similar in pronunciations to the correct (e.g. intended) syllables. We also found that most of the mistakes the subjects made in our Mandarin SUS test were reflected in the list of pairs/groups in the CDRT, CMRT and CDRT-tone. So both the SUS test and the word level tests appear to provide good metrics for intelligibility assessment of highly confusable phonemes.

In the SUS test, however, we also found some incorrectly transcribed syllables that did not form part of the list of the word level tests. For example, syllable 'le' was sometimes wrongly transcribed as 'de', yet the initial pair 'l' – 'd' is not part of the list of the CDRT or CMRT.

Moreover, the CDRT and CMRT do not include syllables without the initial part. Yet, some of these syllables can be easily confusable with others. This confusion cannot be reflected by the current CDRT or CMRT. Also, tone sandhi cannot be tested by isolated hanzi in the CDRT-tone. These shortcomings do not appear in the SUS test.

Finally, in the word level tests, the hanzi lists are fixed which means subjects may be exposed to a learning effect if they do take part in the experiments more than once. In the SUS test, this problem is fixed since hanzis are randomly selected from a dictionary, so a large number of new

sentences can be generated at any one time, thus eliminating the danger of the learning effect

6. Conclusion

This paper attempted to extend the SUS test aimed at intelligibility assessment at sentence level to Mandarin. It was discovered that subjects found the task easier when the sentences were made up of two-hanzi words. However, there was a trade-off between the ease of understanding the syntactic structure and the subjects' short-term memory due to the overall number of hanzis.

One-hanzi words thus appear to be a better choice for the design of the SUS test stimuli in Mandarin. Although they do not present an easier task than two-hanzi words, there is good control for the number of hanzis used in each sentence. Ideally, the number of hanzis used in a sentence should not exceed 7. Further experiments are needed to determine if the syntactic structure contributes to the task when one-hanzi words are used. It was also found that auxiliary words do not contribute to easing the task of remembering sentences.

The comparison of the Mandarin SUS test with CDRT and CMRT revealed that it is useful to include confusable phonemes listed in the CDRT and CMRT in the SUS test. This list is by no means exhaustive, and the results of future SUS test can well be used to augment that list.

7. References

- [1] Z. Li, E. C. Tan, I. McLoughlin and T. T. Teo: Proposal of Standards for Intelligibility Tests of Chinese Speech, IEE Proc., Vis. Image Signal Process., 2000, 147, (3), pp254-260
- [2] Dmitry Sityaev, Katherine Knill and Tina Burrows: Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems, Proc. of ICSLP, 2006, pp 1077-1080
- [3] Christian Benoit, Martine Grice, Valerie Hazan: The SUS Test: A Method for the Assessment of Text-to-speech Synthesis Intelligibility Using Semantically Unpredictable Sentences, Speech Communication, Vol. 18, 1996, pp 381-392
- [4] Christina L. Bennett and Alan W. Black: The Blizzard Challenge 2006, downloaded from the website: <http://www.festvox.org/blizzard/blizzard2006.html>
- [5] Christina L. Bennett: Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005, Interspeech 2005, pp105-108
- [6] Z. Q. Ding, I.V. McLoughlin and E.C. Tan: Extension of proposal of standards for intelligibility tests of Chinese speech: CDRT-tone, IEE Proc., Vis. Image Signal Process., 2003, 150, (1), pp1-5
- [7] Ian McLoughlin, Zhongqiang Ding and Eng Chong Tan: Intelligibility Evaluation of GSM Coder for Mandarin Speech Using CDRT, Speech Communication, Vol. 38, 2002, pp161-165
- [8] Institute of Linguistics, Chinese Academy of Social Sciences: Modern Chinese Dictionary (In Chinese), The Commercial Press, 2002.
- [9] Masatsune Tamura, Tatsuya Mizutani, and Takehiko Kagoshima: Scalable Concatenative Speech Synthesis Based on Plural Unit Selection and Fusion Method, ICASSP2005, pp361-364
- [10] S.J. Young, D. Kershaw, G. Moore, etc.: The HTK Book, downloaded from: <http://htk.eng.cam.ac.uk/>