Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 109-128

Evaluation of TTS Systems in Intelligibility and Comprehension Tasks: a Case Study of HTS-2008 and Multisyn Synthesizers[1]

Yu-Yun Chang[*]

Abstract

This paper explores the relationship between intelligibility and comprehensibility in speech synthesizers, and it designs an appropriate comprehension task for evaluating the speech synthesizers' comprehensibility. Previous studies have predicted that a speech synthesizer with higher intelligibility will have higher performance in comprehension. Also, since the two most popular speech synthesis methods are HMM-based and unit selection, this study tries to compare whether the HTS-2008 (HMM-based) or Multisyn (unit selection) speech synthesizer has better performance in application. Natural speech is applied in the experiment as a control group to the speech synthesizers. The results in the intelligibility test show that natural speech is better than HTS-2008, which, in turn, is much better than the Multisyn system. In the comprehension task, however, all three of the speech systems display minimal differences in the speech comprehension process. This is because the two speech synthesizers have reached the threshold of having enough intelligibility to provide high speech comprehension quality. Therefore, although there is equal comprehensible speech quality between the HTS-2008 and Multisyn systems, the HTS-2008 speech synthesizer is recommended due to its higher intelligibility.

Keywords: Speech Synthesizers, Intelligibility Evaluation, Comprehension Evaluation, HTS-2008, Multisyn.

[1] Acknowledgement: Thanks to the people from University of Edinburgh - for the suggestions and comments offered by Simon King and Catherine Mayo, the perception lab reserving and statistics help supported by Rob Clark, and the HTS-2008 speech synthesizer provided by Junichi Yamagishi.

[*] Graduate Institute of Linguistics, National Taiwan University, 3F, Le-Xue Building, No. 1, Sec. 4, Roosevelt Rd., Taipei Taiwan, 106
E-mail: june06029@gmail.com

1. Introduction

Recently, text-to-speech (TTS) system synthesizers have been evaluated from different aspects, such as intelligibility, naturalness, and preference of the synthetic speech, as noted by Stevens, Lees, Vonwiller, and Burnham (2005). Since the final purpose of applying synthetic speech is to make it usable to applications, carrying out experiments measuring the synthesizers' performance with human listeners is worthwhile.

In previous studies, while mentioning the evaluation of speech synthesizers, most researchers only focused on intelligibility evaluation due to the experiment being easy and quick to carry out. Nevertheless, it is necessary to involve perception factors in synthetic speech evaluation, rather than merely evaluating the intelligibility, in order to better assess speech synthesizers, as indicated by Pisoni, Nusbaum, and Greene (1985). Sydeserff, Caley, Isard, Jack, and Monaghan (1992) also evaluated the aspect of the listener's perception on a comprehension task to learn how well synthetic speech could be understood by the listeners. Moreover, Pisoni *et al*. (1985) demonstrated that intelligibility had a strong impact on comprehension, and specified that intelligibility was one of the important factors affecting listening comprehension. Thus, it is worth observing the linkage between intelligibility and comprehension in speech synthesizers.

Although several studies have evaluated the intelligibility of speech synthesizers successfully, very few researchers have examined its effect on comprehension. This may be because the comprehension measuring experiment is difficult to construct, as it involves cognitive processes that are difficult to capture and take into account. Recent studies have taken post-perceptual comprehension tests instead to investigate listeners' comprehension, but many have failed to distinguish differences between TTS systems. An appropriate strategy for evaluating comprehension still has not been found. Therefore, this research is intended to design an adequate comprehension test for speech synthesis evaluation and to discover the effect of intelligibility on comprehension.

In this study, the word "intelligibility" means the degree of accuracy with which each word is produced in a sentence and the word "comprehension" means the degree of received messages being understood. This study assumes that intelligibility has a strong influence on comprehension, which indicates that speech synthesizers with higher intelligibility can be expected to obtain higher comprehension. In addition, this paper also compares the latest version of speech synthesizers used in the Blizzard Challenge (Black & Tokuda, 2005), which are the unit selection (Clark, Richmond, & King, 2007) based Multisyn synthesizer (Clark, *et al*., 2007) and the hidden Markov models (HMMs) (Zen *et al*., 2007) based HTS-2008 synthesizer (Yamagishi *et al*., 2008). Since these two speech synthesizers are built by adapting the most popular methods used in producing TTS systems, it will be interesting to find out whether the HMM-based or unit selection approach can generate better synthetic speech in terms of both intelligibility and comprehension.

2. Literature Review

2.1 HMM-based and Unit Selection Speech Synthesizers

In recent years, HMMs have been used to generate synthesized speech (Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1999). The basic procedures of implementing HMM-based speech synthesizers to produce synthetic speech can be grouped into two parts: a training part and a synthesis part (Heiga & Tomoki, 2005). There are two main advantages of using HMMs to generate speech synthesizers. One is that the produced synthesized speech can be smoothed and made to sound natural. The other is that, since the synthetic speech is created from HMM models with parameters (Heiga & Tomoki, 2005), the characteristics of the voice can be modified easily with adequate parameter transformations. The latest version of the HTS (HMM-based Speech Synthesis System) used in the Blizzard Challenge is the HTS-2008. HTS-2008 used the speaker adaptive approach, rather than the speaker-dependent method, to generate HMM-based synthesizers. The training database used to create the average voice model for HTS-2008 was a 41-hour speech collection. In addition, to reduce the expensive computing time, the forward-backward algorithm was introduced in HTS-2008 (Yamagishi *et al.*, 2008).

As for the unit selection speech synthesizers, basically, a natural speech database will be recorded by a single speaker and the units are extracted directly from the speech inventory and concatenated together to generate new utterances. A number of different unit sizes can be used to construct various types of unit selection speech synthesizers, such as phones, half phones, diphones, and variable-sized units (Clark, Richmond, & King, 2004). In the recent Festival speech synthesis system, the Multisyn unit selection algorithm was introduced (Clark, *et al.*, 2007) with the diphone sized units, which could carry better acoustic features and higher-level linguistic information than the phone sized units used in CHATR (Hunt & Black, 1996) and clunits (Black & Taylor, 1997). It can produce open-domain speech voices in high speech quality and does not need to be based on the context domain speech to produce better quality. In other words, higher quality synthesized speech can be created using the Multisyn unit selection algorithm even if the synthesized utterance is not one of the sentences in the collected databases.

Since the Multisyn speech synthesis approach has the advantage of generating natural synthesized voices by extracting the diphone sized units straight from the speech signal with less expensive signal processing, an investigation of its distinction from the HTS-2008 HMM-based speech synthesizer would be interesting and useful.

2.2 Evaluation of Intelligibility

When evaluating the intelligibility of a speech synthesizer, semantically unpredictable sentences (SUS) are used frequently. SUS sentences have been widely used in dictation tasks and are recommended in evaluating intelligibility of speech synthesizers (Pols, van Santen, Abe, Kahn, & Keller, 1998). SUS sentences are sentences that are semantically unpredictable, but are still constructed grammatically syntactically. SUS sentences are used to prevent the process of assessing intelligibility from being influenced by linguistic cues. If semantically predictable sentences are used, listeners will learn the semantic and syntactic cues from the context, which will influence their performance in the intelligibility task (Benoît, Grice, & Hazan, 1996). They claimed that using SUS sentences in the intelligibility task could disrupt the predictable context. This conclusion was also supported by Miller and Isard (1963), reporting that using SUS sentences could prevent the learning effect.

2.3 Evaluation of Comprehension

The performance of various speech synthesizers can also be evaluated through comprehension tasks. Several researchers have indicated that comprehension evaluation is a valid way to assess intelligibility (Hustad, 2008; Yorkston, Strand, & Kennedy, 1996). This is because, in the intelligibility task, listeners will emphasize recognizing individual words, rather than focusing on the meaning of sentences. Nevertheless, the deeper information that lies within intelligibility cannot be examined by merely identifying each word.

There are four types of questions that have been used in speech synthesizer comprehension evaluation: surface structure questions, high proposition questions, low proposition questions, and inference questions. These questions were designed based on different levels of memory used during comprehension (Luce, 1981; Pisoni, Nusbaum, Luce, & Schwab, 1983; Salasoo, 1982). Surface structure questions required participants to recall specific words that occurred in the speech content. High proposition questions examined whether listeners could get a general idea from the speech content, whereas low proposition questions asked for more detailed information about the speech content than high proposition questions. Finally, the inference questions measured whether the listeners could draw a conclusion from the speech. Since surface structure questions did not involve much comprehension ability, which did not meet the purpose of the present experiment, this type of question was not included in the present study.

2.4 Some Influential Factors in Intelligibility and Comprehension

2.4.1 Short-term Memory

Short-term memory is the biggest cognitive factor influencing the comprehension task. This is because short-term memory is used to store fractions of information temporarily until full information can be completely comprehended. Therefore, the technique is essential during the comprehension task, and the load of short-term memory needs to be considered as well. As demonstrated from the concurrent task experiment by Ralston, Pisoni, and Mullennix (1989), short-term memory has limited capacity. Goldstein (1995) identified two different levels of short-term memory, which are the nominal level and supra-nominal level. He further said that nominal level short-term memory was involved in intelligibility tasks, focusing on qualitative evaluation, whereas supra-nominal level short-term memory was used in comprehension tasks, which required the information to be identified, processed, and understood. Therefore, as specified by previous researchers, it would be important to take short-term memory into account in this study.

2.4.2 Listeners' Preferences

Another factor that may influence task performance is the listeners' preferences. Nusbaum *et al*. (1984) judged listeners' preferences from listeners' feedback on one natural speech and two speech synthesizers, MITalk and Votrax. The measurement was to assess adjectives from the feedback. The researchers found that, although people preferred to listen to natural speech rather than the two speech synthesizers, they liked the MITalk system more than the Votrax system. Also, they investigated the intelligibility in the MITalk system and evaluated it as higher than the Votrax system. As indicated in the paper, this result showed that a relationship existed between the subjects' preferences and intelligibility of different speech synthesizers. Besides, Nusbaum, Francis, and Henly (1995) contended that listeners' preferences depended greatly on the quality of speech intelligibility. Moreover, Terken and Lemeer (1988) and Paris, Thomas, Gilson, and Kincaid (2000) found that, as the intelligibility got better, the degree of preference would also increase.

Therefore, in this paper, HTS-2008 and Multisyn systems would be taken as the representatives of HMM-based and unit selection speech synthesizers during the evaluation. Also, by modifying the evaluation approaches used in the previous studies and considering the cognitive factors, I try to design an appropriate comprehension test, which has not been found yet, rather than designing an intelligibility test. In addition, through the newly modified comprehension test, I hope that a stronger relationship of "higher intelligibility will gain better comprehension" could be revealed.

3. Methodology

3.1 Subjects

Twenty-five native English speakers participated in the experiment, with 6 males and 19 females.[2] Table 1 shows the subjects' level of education.

Table 1. Participants' level of education status

Degree of Education	Undergraduate	Master	PhD
Number of Subjects	5	11	9

All of the participants were students studying at University of Edinburgh at the time of the survey. There were 5 undergraduates, 11 master's students, and 9 PhD students involved in this experiment. The subjects' average age was 25.44 years old, with a standard deviation (SD) of 3.465 years.

Table 2. Participants' English accents

English Accent	British	American	Scottish	Irish	Welsh	Indian
Number of Subjects	13	6	3	1	1	1

Table 2 presents the survey results of the participants' English accents. The accent survey reported 13 people with a British accent, 6 with an American accent, 3 with a Scottish accent, 1 with an Irish accent, 1 with a Welsh accent, and 1 with an Indian accent. Additionally, only three participants indicated that they were speech experts. No one reported having a hearing disorder.

3.2 Materials

3.2.1 SUS Sentences for Intelligibility Evaluation

Thirty SUS sentences were used as the material in the intelligibility task. These SUS sentences were adopted from the 2008 Blizzard Challenge (Karaiskos, King, Clark, & Mayo, 2008). The structure of these sentences is "The (Determiner) + (Adjective) + (Noun) $_{\text{plural}}$ + (Verb) $_{\text{past tense}}$ + the (Determiner) + (Adjective) + (Noun) $_{\text{singular}}$". Although this was the only structure used in the experiment, the English words chosen to construct SUS sentences are all low-frequency words, in order to prevent the listeners from predicting meanings easily. For example, one of the sentences used in the experiment is "The amicable chests became the unprepared cockroach". As the example shows, the intelligibility task tends to make it difficult for

[2] Although the numbers of male and female participants were not balanced, the gender did not display any significance in statistical analysis. Therefore, the gender difference is not considered in this paper.

listeners to predict the unheard information. In addition, listening to each sentence more than once was allowed, but subjects were requested to keep this to as few times as possible.

3.2.2 News Articles for Comprehension Evaluation

Six news articles from BBC News online that were considered to contain few story line cues were used in the comprehension task. As in the study of Lai, Wood, and Considine (2000), in order to reduce the news articles' textual familiarity to the listeners, all of the topics chosen were research reports, which were likely to be less familiar to most of the listeners. The answers to the questions were designed with the assumption that there was no global and general knowledge to the articles. In other words, participants could not learn the answers to questions without listening. The average article was about 238.8 words (SD = 21.1 words).

Each news article was attached to ten questions. Five of the questions were designed as multiple-choice questions, while the other five questions were open-ended questions. Only the questions that required inferential skills would be arranged as multiple-choice questions with four choices. On the other hand, factual questions with low-level proposition information were assigned to open-ended questions. Figures 1 and 2 present examples of the questions involved in the main experiment.

<u>Inferential Question</u>

Question: What would be the best topic for the news?

A. The poor quality of recent education.
B. The cpmpetition between colleges.
C. Colleges face the financial crisis.
D. Education revolution.

Figure 1. An example of inferential question in the main experiment

<u>Factual Question</u>

Question: How long would the growth of stubble usually appears?

____

Figure 2. An example of factual question in the main experiment

3.2.3 Synthesized Speech and Natural Speech Recording

HTS-2008 and Multisyn speech synthesizers were included in this experiment. Both speech synthesizers were constructed by collecting the voice from a single male speaker with a British accent, "Roger". Also, the male speaker's natural speech was taken as a control group, to compare with the experimental materials (30 SUS sentences and 6 news articles) produced by the two synthesizers.

The recording was held in a sound lab of University of Edinburgh. The lab was equipped with a professional recording room and a control room. The voice was recorded through a Sennheiser MKH 800 microphone, with the volume set at 60 dB. The recorded wav files were all single channel, with a frequency of 16 kHz. The recording duration was approximately one hour.

The male speaker was a well-trained professional reader and had cooperated with the Centre for Speech Technology Research (CSTR) for a long while, participating in speech data recording. Therefore, steady and good quality natural speech was guaranteed.

3.2.4 Questionnaires

A questionnaire was assigned at the end of the experiment, asking for participants' basic information, whether they were speech experts, and the average number of times each sentence in the intelligibility task was played. Some empty blanks were left for participants to write down their comments and suggestions about the experiment.

3.3 Procedure

There were two tasks in the experiment. The first part was an intelligibility task (listening to 30 SUS sentences), and the other part was the comprehension task (listening to 6 BBC News reports and answering questions). The experiment took place at the Perception Lab in the Informatics Forum building. The lab consisted of individual rooms. Each room was equipped with a SAMSUNG 2043 screen monitor and a set of Beyerdynamic DT 770 PRO headphones. Every participant was arranged into one of the single rooms. The experiment was carried out by applying an online webpage. All of the voices would come from the headphones throughout the experiment, and the volume had been set to an adequate loudness for the listeners. No participants complained about the sound volume.

3.3.1 Producing Wav Files

For the intelligibility task and comprehension task, all wav files of SUS sentences and news passages were produced by natural speech and the two synthesizers, HTS-2008 and Multisyn. In order to generate higher-quality synthesized speech for news passages, all of the sentences in each article were synthesized individually before being concatenated together with a silence interval of about 500 milliseconds in between.

There were some cases needing careful consideration when producing synthesized speech, where the TTS systems could not identify the pronunciation as predicted in natural speech. For example, if the input text was “500MB,” the synthesizers would not be able to pronounce it as “five hundred megabytes”. Instead, the pronunciation turned out to be “five zero zero M B”. Since the purpose of this comprehension test was to measure whether the synthesized passages were comprehensible to listeners, every word in the experiment should be made understandable to listeners.

3.3.2 Pilot Tests for Comprehension Task

Since the material used in the intelligibility test was the same as in the Blizzard Challenge, pilot tests for evaluating the intelligibility test were unnecessary. Nevertheless, pilot tests were needed for the comprehension test in this study. The pilot tests for the comprehension test were done three times, measuring the length of the articles, the difficulty of the text and questions, and the familiarity of the text. Two native English speakers were invited to do the pilot test and help evaluate the design of the comprehension task.

3.3.3 Main Experiment

To make the wav files produced from HTS-2008, Multisyn, and natural speech equally distributed through the experiment, the wav files were equally arranged into 6 different groups via Latin Squares. Each group included 30 SUS sentences in the intelligibility test and 6 news articles in the comprehension test. Then, each listener would be assigned to one of the six groups. Also, in order to prevent the participants from having pressure taking the exams, an announcement was made beforehand indicating that they were testing the systems, not being tested.

The intelligibility task was taken before the comprehension task. It was arranged this way due to more effort being required in the comprehension task than in the intelligibility test, where participants needed to answer questions rather than simply type the words they heard. Therefore, it would be better not to depress the listeners’ patience and willingness in the first task. The listeners were informed in advance that the sentences in the intelligibility task might not be meaningful to them and were requested to try to listen as few times as possible. For the comprehension task, listeners were only allowed to listen to each news article once before answering questions without taking notes. Also, two extra subjective questions followed each news article, asking about the participants’ confidence in completing the questions and their feelings about the speech quality, scaled from 1 (very low) to 5 (extremely high). Finally, a questionnaire was given after completing the two tasks.

The intelligibility task of this experiment took around 15 to 20 minutes, while the comprehension test was about 25 to 30 minutes. Delogu *et al.* (1998) pointed out that many researchers had found that participants would fail to maintain their attention after 20 to 35 minutes of doing the task. Due to this finding, participants were asked to take a 5-minute break between the two tasks.

4. Results[3]

4.1 Intelligibility Task

Most of the participants specified that they only listened to each sentence once, and typed what they heard. For assessing SUS sentences, the measurement was based on calculating word error rates (WER) occurring in every sentence. Typographical errors and homophones were allowed.

Table 3. Significant differences in intelligibility of the three speech systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems.[4]

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		■
Multisyn	■	■	

In Pairwise Comparisons, as presented in Table 3, there are significant differences found between natural speech and HTS-2008 ($p = 0.005$), natural speech and Multisyn ($p < 0.001$), and HTS-2008 and Multisyn systems ($p < 0.001$). To further verify the main effects in Pairwise Comparisons, the results in the Tests of Within-Subjects Contrasts show that there are significant effects when natural speech is compared to HTS-2008, $F(1, 249) = 10.135$[5], $p = 0.002$; and when HTS-2008 is compared to the Multisyn system, $F(1, 249) = 26.685$, $p < 0.001$. Therefore, it can be concluded that natural speech has significantly lower WER (M = 4.2%, SD = 10%) than HTS-2008 (M = 6.7%, SD = 11.4%) and HTS-2008 is even better than the

[3] Since a detailed table of the scored collected from intelligibility and comprehension tests might be too much to confuse the results description in this section, I simply provide tables with further analysed statistical results here.

[4] There are a total of 4 figures in this paper describing the statistical significant differences between speech synthesizers based on experimental results. Combined with the results presented in the figures, the statistical mean value (M) and standard deviation value (SD) are also given to further investigate their performance.

[5] In this section, you will find that a lot of statistical values are provided. In the presented form, $F(a, b) = c$, F is the symbol of degree of freedom (df); a is the df value in the whole tested data set; b is the df value of the deviation between the data set; and c is the output value of df. When the distance between a and b values gets larger, the greater the c value represents a stronger significant difference existed within the data set, usually followed with a p value as a reference.

Multisyn system (M = 14.3%, SD = 21.6%).

4.2 Comprehension Task

4.2.1 The Results from News Articles

A 3-point scale (0, 1, 2) was applied in the experiment to score answers in the open-ended questions. If the responses to the comprehension questions were judged to be incorrect, 0 points were earned; if part of the answers were correct or the answers were too general and nonspecific, yet not wrong, 1 point would be given; and 2 points were given to the responses with fully correct and specific answers. A total of 10 points for 5 open-ended questions per news article was possible. The examples of assessing the responses from open-ended questions are provided in Table 4.

Table 4. Examples of assessing the responses from open-ended questions

Open-ended Question	Correct Answer	Listener Response	Score
What are the two new news channels that have been launched by Russia?	English and Arabic	English, Arabic	2
		English and Polish	1
		Arabic	1
		Don't know	0

The 3-point scoring system was adopted from Hustad (2008). The reason for not taking a 2-point binomial scoring scale was because, in real life comprehension, it is not always an all correct or wrong situation, as described by Hustad & Beukelman (2002). Nevertheless, since the multiple-choice questions only had one correct answer, the binomial scoring system was introduced to assess the responses. If the participants chose the correct choice, then 2 points would be earned; if they chose the wrong answer, 0 points would be awarded. There would be a sum of 10 points for 5 multiple-choice questions per news article. Therefore, the total score in each article was 20 points.

There is no significance found in the three speech systems and none in the interaction between systems and the question types. Nevertheless, there is a significant effect occurring in the question types, $F(1, 24) = 29.004$, $p < 0.001$. Therefore, the performance in open-ended questions was considerably worse (mean of error rate = 39.1%) than multiple-choice questions (mean of error rate = 28%). Furthermore, there is no significance found in the interactions between the systems and multiple-choice questions. Nevertheless, there is a main effect observed in the interaction between systems and open-ended questions, $F(1.569, 37.649) = 7.348$, $p = 0.004$. Due to this fact, it can be interpreted that the results from open-ended questions shows the differences of the three systems.

Table 5. Significant differences in open-ended questions of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural			
HTS-2008			■
Multisyn		■	

As presented in Table 5, in the open-ended questions, a significant effect is revealed only when the comparison is between HTS-2008 and the Multisyn system, $F(1, 24) = 25.939$, $p < 0.001$. Also, HTS-2008 performs much better (mean of error rate = 29.2%) than the Multisyn system (mean of error rate = 49.8%) in answering the open-ended questions correctly.

4.2.2 A 5-point Scale for Subjective Judgments

Two individual subjective questions were given at the end of each news articles: the confidence in making right responses to the questions (Confidence) and the feeling about the displayed speech quality (Quality). Both the Confidence and Quality tests used a 5-point scale (from 1 to 5) in assessing the subjective questions. Higher points represented listeners with higher satisfaction, as shown below in Table 6.

Table 6. The 5-point scale measurement for the Confidence and Quality subjective tests

1 = Very low.
2 = Low.
3 = Average
4 = High.
5 = Extremely high.

Accordingly, there are main effects found in the systems, $F(1.45, 34.806) = 25.365$, $p < 0.001$, and in the interaction between systems and the subjective tests, $F(2, 48) = 58.808$, $p < 0.001$. Nevertheless, there is no significant main effect observed in the subjective tests.

Table 7. Significant differences in the overall subjective test performance of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		
Multisyn	■		

In Table 7, highly significant effects occurred when HTS-2008 was compared to natural speech, $F(1, 24) = 24.758$, $p < 0.001$; and when the Multisyn system was compared to natural speech, $F(1, 24) = 37.536$, $p < 0.001$. While Quality compares to Confidence, two main effects are discovered in the interactions when HTS-2008 is compared to natural speech, $F(1, 24) = 89.161$, $p < 0.001$, and when Multisyn is compared with natural speech, $F(1, 24) = 73.059$, $p < 0.001$. Therefore, it can be concluded that HTS-2008 is evaluated lower (M = 52.4%) than natural speech (M = 71.6%) in the subjective tests and lower points are given to Multisyn (M = 52.2%) than to natural speech. Therefore, it is known that natural speech has better results from the subjective tests than the HTS-2008 and Multisyn systems.

The Confidence test does not show any significant effect on the systems. This result indicates that listeners have equal confidence in natural speech, HTS-2008, and the Multisyn system in answering the questions of each news article. As for the results from the Quality test, there is a significance discovered in the systems, $F(1.462, 35.085) = 61.249$, $p < 0.001$.

Table 8. Significant differences in Quality test of the three systems: results of Pairwise Comparisons. ■ indicates a significant difference between a pair of systems

	Natural	HTS-2008	Multisyn
Natural		■	■
HTS-2008	■		
Multisyn	■		

In the Quality test, natural speech has an extremely high score in speech quality identification (M = 82.8%), compared to the HTS-2008 (M = 48.8%) and Multisyn (M = 49.6%) systems. The results in Table 8 show no significance when HTS-2008 is compared to the Multisyn system. As a result, in the subjective judgment of speech quality, natural speech is scored significantly higher than HTS-2008 and Multisyn. On the other hand, the HTS-2008 and Multisyn systems are rated with nearly the same synthetic speech quality by listeners.

The results also demonstrate that, although all of the news articles were generated by concatenating the individual sentences together, natural speech still has better speech prosody than the other two speech synthesizers. This is because the recorder of natural speech knows the context and will be able to articulate the sentences with adequate prosody contours while recording. Nevertheless, the news articles produced by HTS-2008 and Multisyn systems were simply synthesized into individual sentences, without considering the context prosody factor. As stated by Sanderman and Collier (1997), listeners preferred the speech systems with higher prosody quality. Therefore, the listeners graded natural speech with the highest score, compared to HTS-2008 and Multisyn.

5. Discussion

5.1 The Discussion in the Experiment Results

5.1.1 The Relationships between Intelligibility and Comprehension

In the intelligibility task, the results prove there are significant differences between the three systems. In the intelligibility performance, natural speech is better than HTS-2008, while HTS-2008 has greater performance than the Multisyn system. According to the initial assumption in this paper, assuming systems with higher achievement in the intelligibility task would also preserve better accomplishment in the comprehension task, we can estimate that the three systems might have the same rankings in the comprehension task as presented in the intelligibility task.

Nevertheless, in the overall comprehension task performance, no significant effects are noticed within the three systems, which signifies that natural speech, HTS-2008, and Multisyn all have a similar understandability quality for listeners. The outcomes in the comprehension task are against the results in the intelligibility task and violate the assumptions of this paper. Although it seems that the comprehension task in this study has also failed to distinguish various speech systems, this is mainly because the three systems have reached the threshold of producing comprehensible speech quality. This can be demonstrated from the results in the Confidence test.

In the Confidence test, there was no significant difference observed in the three systems, which meant that listeners have equivalent confidence in completing the comprehension task produced by the systems. This implies that the three systems have given identical comprehension quality to the listeners. In addition, the techniques required for evaluating intelligibility and comprehension are different.

In the comprehension task, the main intention is to understand and comprehend the global meanings offered in each news article, whereas the intelligibility task is not evaluated by focusing on the meanings of the words but on paying attention to every single word that can be heard. During the process of comprehension, even if some of the words are not clear to the listeners, the comprehension process will not be interrupted. Listeners can still acquire general meanings from the context of the articles. Benoît *et al*. (1996) found that, with sufficient linguistic cues, it would be easy for listeners to derive learning effects and process the effects while comprehending. Thus, with sufficient cues provided from the three systems, no significant differences could be found within the three systems in the comprehension task. In other words, although natural speech, HTS-2008, and Multisyn are significantly different from each other in intelligibility, they all obtain enough intelligibility quality for listeners to learn the linguistic cues and comprehend the texts. In addition, the WER of 14.3% in the Multisyn system can be taken as an intelligibility threshold reference for achieving high comprehensibility in speech synthesizers.

5.1.2 The Influence of Different Question Types used in the Comprehension Task

In the comprehension task, different question types used in the experiment will bring a significant effect to the systems' measurement. In this experiment, only the open-ended questions have a significant effect on the systems. This may be affected by the design purpose of each type of question.

For the multiple-choice questions, they are assigned to be inferential questions, which need to be processed and comprehended before answering. Thus, this procedure is very much the same as in the real comprehension process and shows that natural speech, HTS-2008, and Multisyn have the same comprehensibility. Nevertheless, the open-ended questions are designed to be factual questions, which makes the process of answering the questions similar to the way of completing the intelligibility task. Both the open-ended questions and intelligibility task involve listening to the speech first and focusing on the key words they can capture or understand.

The only difference between them is that the load of memory will be larger in open-ended questions than in the intelligibility task. As seen in the results of open-ended questions, the consequences diverge a little from the results in the intelligibility task. In the open-ended questions, the performance in natural speech is identical to HTS-2008, but is better than the Multisyn system. The intelligibility task, however, shows that natural speech is better than HTS-2008 and Multisyn. In addition, even the overall subjective tests and quality test show that natural speech has better achievement than HTS-2008. This may be due to there not being enough participants included in the experiment (only 25 participants in this study).

Therefore, it is assumed that, if the number of participants increases, the significant effect between natural speech and HTS-2008 in open-ended questions might occur. Apart from the intelligibility and comprehension task, the overall subjective tests and quality test are both consistent with the results specifying that the performances in HTS-2008 and the Multisyn system are the same. In general, the entire experiment in the present study has found that natural speech has greater impact and performance than HTS-2008 and Multisyn.

5.2 Listeners' Feedback and some Suggestions for Future Studies

5.2.1 Listeners' Feedback

Most of the participants found the intelligibility task interesting. Since the materials were all semantically unpredictable sentences, there could be many unexpectedly funny sentences. Still, some of the participants specified that there were a few words they had seldom heard or seen in their life, which might lead to some misspelling or making up the spelling. This problem was solved in this study by allowing typographical errors and homophones while calculating the WER in the intelligibility task. They also indicated that, in sentences with poor speech quality, it would be difficult for them to recognize the words as real words.

Most of the participants reported that the second part of the experiment (comprehension task) was harder than the first part (intelligibility task). They stated that the display duration of news articles was a bit long for them to remember all of the information. Besides, the listeners stated that, if the article were presented with low speech quality, it would be harder for them to concentrate and follow up. In addition, they tended to focus more on the topics they were interested in and answered these questions correctly more often. Some participants suggested that there should be an option of "do not know the answer" added to the multiple-choice questions to prevent them from guessing the answers.

Although there were comments coming from the participants, they still responded that the whole experiment was interesting, and they had a lot of fun during the process.

5.2.2 Suggestions and Modifications for Future Works

According to the feedback received from the participants, some things can be modified in the comprehension design to make the task better. First, since most of the participants replied that the durations of news articles were a little bit too long, a pilot test for measuring the participants' feelings of duration needs to be applied before carrying out the main experiment. In addition, with long news articles as experimental materials, there may be too much redundant information embedded, which may interfere with the comprehension testing. Furthermore, since each news article had different topics, there is no guarantee that the degree of text complexity and familiarity would remain the same between articles. The word "text complexity" used here means the degree of comprehension effort that needs to be devoted to listening to the article.

Due to the limitation of time, there were not enough listeners participating in each pilot test. In order to remove the individual problems and increase the objectivity of the results of the test, it will be better to have at least 10 people included in the pilot test.

6. Conclusion

From the results in the intelligibility task, we find that the performance in natural speech is better than HTS-2008, and HTS-2008 is proven better than the Multisyn system. Nevertheless, the results in the comprehension task show that the natural speech, HTS-2008, and Multisyn systems display equal quality for listeners to comprehend. The explanation has been given in Section 5.1.1, discussing the issue that all three systems obtain enough intelligibility quality to be used in comprehending the news passages. Although the outcomes in the intelligibility task show that there are significant differences in the three systems, their intelligibility has reached the comprehension threshold to produce understandable high quality speech. In spite of the objective results in the comprehension task, in the overall subjective tests and the Quality test, both of them show that listeners consider natural speech to be the best system of all, compared to the two speech synthesizers (HTS-2008 and Multisyn). Besides, the listeners feel that there is no difference between HTS-2008 and the Multisyn system.

For the design of the comprehension task, there is still one thing that needs to be mentioned. That is the comprehension task designed in this experiment could not directly evaluate the comprehension process, as stated by Pisoni *et al*. (1985). Since the questions are derived after listening, this kind of measurement is a post-perceptual comprehension. Therefore, the comprehension strategies involved in this study are all evaluating the products of comprehension, rather than the process itself.

In general, from the results of this experiment, the HTS-2008 speech synthesizer is preferable and more usable than the Multisyn system. Although the two systems have the same performance in comprehension, HTS-2008 is significantly better than the Multisyn system in intelligibility.

References

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication,* 18, 381-392.

Black, A. W., & Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of the Eurospeech 1997.*

Black, A. W., & Tokuda, K. (2005). The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common dataset. In *Proceedings of Interspeech 2005,* Lisbon, Portugal.

Clark, R. A. J., Richmond, K., & King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesiser. In *Proceedings of 5th ISCA Speech Synthesis Workshop,* Pittsburgh, USA.

Clark, R. A. J., Richmond, K., & King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication,* 49, 317-330.

Delogu, C., Conte, S., & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication,* 24, 153-168.

Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication,* 16(3), 225-244.

Heiga, Z., & Tomoki, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proceedings of Interspeech 2005,* Lisbon, Portugal.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the ICASSP 1996,* Atlanta, USA.

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research,* 51, 562-573.

Hustad, K. C., & Beukelman, D. R. (2002). Listener coomprehension of severely dysarthric speech: Effects of linguistic cues and stimulus cohesion. *Journal of Speech, Language, and Hearing Research,* 45, 545-558.

Karaiskos, V., King, S., Clark, R. A. J., & Mayo, C. (2008). The Blizzard Challenge 2008. In *Proceedings of the Blizzard Challenge 2008 workshop*, Brisbane, Australia.

Lai, J., Wood, D., & Considine, M. (2000). *The effect of task conditions on the comprehensibility of synthetic speech*. Paper presented at the CHI Letters.

Luce, P. A. (1981). *Comprehension of fluent synthetic speech produced by rule* (Research on Speech Perception Progress Report No. 7). Bloomington, IN 47405: Indiana University.

Miller, G. A., & Isard, S. D. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior,* 2, 217-228.

Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology,* 1, 7-19.

Nusbaum, H. C., Schwab, E. C., & Pisoni, D. B. (1984). *Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability* (Research on Speech Perception Progress Report No. 10). Bloomington, IN47405: Speech Research Laboratory, Indiana University.

Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors,* 42, 421-431.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. In *Proceedings of the IEEE*.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Schwab, E. C. (1983, Apr 1983). Perceptual evaluation of synthetic speech: Some considerations of the user/System interface. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83*.

Pols, L. C. W., van Santen, J. P. H., Abe, M., Kahn, D., & Keller, E. (1998). The use of large text corpora for evaluation text-to-speech systems. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Ralston, J. V., Pisoni, D. B., & Mullennix, J. W. (1989). *Comprehension of synthetic speech produced by rule* (Research on Speech Perception Progress Report No. 15). Bloomington, IN47405: Speech Research Laboratory, Indiana University.

Salasoo, A. (1982). *Cognitive Processes and comprehension measures in silent and oral reading* (Research on Speech Perception Progress Report No. 8). Bloomingtion, IN 47405: Speech Research Laboratory, Indiana University.

Sanderman, A. A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech,* 40, 391-409.

Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language,* 19, 129-146.

Sydeserff, H. A., Caley, R. J., Isard, S. D., Jack, M. A., & Monaghan, A. I. C. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech Communication,* 11, 189-194.

Terken, J., & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics,* 16, 453-457.

Yamagishi J., *et al.* (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008,* Brisbane, Australia.

Yorkston, K., Strand, E., & Kennedy, M. (1996). Comoprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology,* 5(1), 55-66.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultanious modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of Eurospeech 1999*.

Yu, S.-Z., & Kobayashi, T. (2003). An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Processing Letters,* 10, 11-14.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., et al. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of ISCA SSW6,* Bonn, Germany.