Evaluation of TTS Systems in Intelligibility and Comprehension Tasks[1]

張瑜芸　Yu-Yun Chang
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
june06029@gmail.com

Abstract

This paper aims at finding the relationships between intelligibility and comprehensibility in speech synthesizers, and tries to design an appropriate comprehension task for evaluating the speech synthesizers' comprehensibility. It is predicted that speech synthesizer with higher intelligibility, will have greater performance in comprehension. Also, since the two most popular used 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 controlled group to the speech synthesizers. The results in the intelligibility test shows that natural speech is better than HTS-2008, and HTS-2008 is much better than Multisyn system. Whereas, in the comprehension task, all the three speech systems present not much differences in speech comprehending process. This is because that the two speech synthesizers have reached the threshold of enough intelligibility to provide high speech comprehension quality. Therefore, although with equal comprehensible speech quality between HTS-2008 and Multisyn systems, HTS-2008 speech synthesizer is more recommended and preferable due to its higher intelligibility.

Keywords: speech synthesizers, intelligibility evaluation, comprehension evaluation, HTS-2008, Multisyn

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 [1]. Since the final purpose of applying the synthetic speech is to make it usable to applications, it is worth carrying out experiments measuring the synthesizers' performance with human listeners. For measuring speech synthesizers, it was necessary to involve perception factors in synthetic speech evaluation, rather than merely evaluating the intelligibility, in order to better assess the speech synthesizers, as indicated by [2]. [3] also evaluated the aspect of the listener's perception on a comprehension task to learn how well the synthetic speech was created by the synthesizers could be understood by the listeners. Moreover, [2] had demonstrated that there was a strong relationship between the intelligibility and comprehension. Also, they had specified the intelligibility was one of the important factors that would affect listening comprehension. It is then worth observing the relationships between intelligibility and comprehension for speech synthesizers. Although several studies have been successfully evaluating the intelligibility of speech synthesizers, very few researchers have examined the association with comprehension. However, it is hard

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

to measure comprehension, due to the fact that it involves cognitive processes which are hard to be captured and taken into account. Recent studies use post-perceptual comprehension tests to measure listeners' comprehension, but many have failed to distinguish differences between TTS systems. An appropriate strategy for evaluating the comprehension is still not found. Therefore, this research aims to design an adequate comprehension test for speech synthesis evaluation, and to try to discover the relationship between intelligibility and comprehension of TTS systems. In this study, the word "intelligibility" means the degree of each word being produced in a sentence; while the word "comprehension" means the degree of received messages being understood. This study predicts that speech synthesizers with higher intelligibility can be expected to obtain higher comprehension. In addition, this paper will also compare the most popular methods for building TTS systems in the Blizzard Challenge [4], which are unit selection [5] and hidden Markov models (HMMs) [6]. It will be interesting to find out whether the HMM-based, or unit selection approaches can better generate 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 [7]. 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 [8]. 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 [8], the characteristics of the voice can be modified easily with adequate parameter transformations. Nowadays, the latest version of the HTS (HMM-based Speech Synthesis System) used in the Blizzard Challenge is the HTS-2008. HTS-2008 used the adaptive speaker-independent approach, rather than the speaker-dependent method, to generate HMM-based synthesizers. The training database for HTS-2008 using the average voice model was 41 hours. In addition, to reduce the expensive computing time, forward-backward algorithm was introduced in the HTS-2008 [9].

As for unit selection speech synthesizers, basically, a natural speech database will be recorded by a single speaker, and then 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 [10]. In recent Festival speech synthesis system, the Multisyn unit selection algorithm was introduced [5] with the diphone sized units, which could carry better acoustic features and higher level linguistic information than the phone sized units used in CHATR [11] and clunits [12]. 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 by using Multisyn unit selection algorithm even if the synthesized utterance is not one of the sentences in the recorded 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 distinctions from the HTS-2008 HMM-based speech synthesizer will be interesting and useful.

2.2 Evaluation of intelligibility

When evaluating the intelligibility of a speech synthesizer, the semantically unpredictable sentences (SUS) are frequently used. SUS sentences have been widely used in a dictation task and are recommended in evaluating intelligibility of speech synthesizers [13]. 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 [14]. [14] claimed that using SUS sentences in the intelligibility task could disrupt the predictable context. This conclusion was also supported by [15] reported that using SUS sentences could prevent from learning effect.

2.3 Evaluation of comprehension

The performance of various speech synthesizers can also be evaluated through comprehension tasks. Several researchers had indicated that comprehension evaluation is a valid way to assess intelligibility [16, 17]. This is because in intelligibility task, listeners will emphasize on recognizing individual words, rather than focusing on the meaning of sentences. However, the deeper information that lies within intelligibility cannot be examined by merely identifying each word.

There were four types of questions that had been used in previous 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 [18-20]. 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; low proposition questions asked 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 with the purpose of present experiment, this type of question was not included in present study.

2.4 Some influential factors in intelligibility and comprehension

2.4.1 Short-term memory

The short-term memory is the biggest cognitive factor that has the greatest influence on 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 quite essential during the comprehension task. Furthermore, the load of short-term memory needs to be considered as well. As demonstrated from the concurrent task experiment by [21], the short-term memory had limited capacity. Goldstein [22] had identified two different levels of short-term memory, which were nominal level and supra-nominal level. He described that the nominal level short-term memory was involved in intelligibility tasks, focusing on qualitative evaluation. On the other hand, the supra-nominal level short-term memory were 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. [23] judged listeners' preferences from listeners' feedback on one natural speech and two speech synthesizers: MITalk and Votrax. The measurement was to assess the adjective words from the feedback. The researchers found that people preferred to listen to natural speech than to the two speech synthesizers, and MITalk system was preferred than Votrax system. Also, the intelligibility in MITalk system was evaluated to be higher than Votrax system. This result presented that there was a relationship between subjects' preferences and intelligibility of different speech synthesizers. Besides, [24] contended that listeners' preferences depended greatly on the quality of speech intelligibility. Moreover, [25] and [26] investigated that as the intelligibility quality 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 some cognitive factors, I try to design an appropriate comprehension test, which has not been found yet, rather than intelligibility test. In addition, through the newly modified comprehension test, I hope that stronger relationships between intelligibility and comprehension could be revealed.

3. Methodology

3.1 Subjects

A total of 25 native English speakers participate in the experiment, with 6 male and 19 female[2]. Table 1 shows the subjects' highest level of education status.

Table 1. Participants' highest level of education status

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

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

Table 2. Participants' English accents

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

Table 2 above presents the survey results of the participants' English accents. In the English accent survey, 13 people have reported that they have a British accent, 6 have an American accent, 3 have a Scottish accent, 1 has an Irish accent, 1 has a Welsh accent, and 1 has an Indian accent. Only three participants have indicated that they are speech experts. No one has reported having a hearing disorder.

3.2 Materials

3.2.1 SUS sentences for intelligibility evaluation

Thirty SUS sentences are used as the material in intelligibility task. These SUS sentences are adopted from the 2008 Blizzard Challenge [27]. The structure of these sentences is "The (Determiner) + (Adjective) + (Noun) $_{\text{plural}}$ + (Verb) $_{\text{past tense}}$ + the (Determiner) + (Adjective) + (Noun) $_{\text{singular}}$". Although, this is the only structure used in the experiment, the English words in the SUS sentences are all in low frequency, in order to prevent the listeners from predicting the meanings easily. For example, one of the sentence used in the experiment is "The amicable chests became the unprepared cockroach". As the example shows, the intelligibility

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

task tends to make listeners hard to foretell the unheard information. In addition, listening to each sentence more than once is allowed, but are requested to keep as few times as possible.

3.2.2 News articles for comprehension evaluation

6 news articles extracted from BCC online news, which were considered to contain less story line cues, were used in the comprehension task. As in the study of [28], in order to reduce the news articles' text familiarity to the listeners, all of the topics were chosen to be 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 were no global and general knowledge to the articles. In other words, participants could not learn the answers through questions without listening. The average words in each article was about 238.8 words (SD = 21.1 words).

Each news article was attached with 10 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 4 multiple choices. On the other hand, factual questions with low level proposition information were assigned to open-ended questions. Below are figure 1 and 2, presenting the examples of the questions involved in the main experiment.

Inferential Question

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

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

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

Factual Question

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

____

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 "roger" with British accent. Also, the male speaker's natural speech was taken as a controlled group in the experiment, to compare with 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 the MKH800 microphone, with the volume set at 60 dB. The recording wav files were all in single channel, with frequency at 16 kHz. The whole recording duration lasted approximately an hour.

The male speaker was a well-trained and professional reader, and had been cooperated with the Center for Speech Technology Research (CSTR) for a long while, participating in speech data recording. Therefore, steady and good quality of the 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 a speech expert, and the average playing times of each sentence in intelligibility task. Some empty blanks were left for participants to write down their comments and suggestions to the experiment.

3.3 Procedure

There were two tasks included in the experiment. The first part was intelligibility task (listening 30 SUS sentences), and the other part was the comprehension task (listening 6 BBC news articles and answering questions). The experiment was taken place at the Perception Lab within the Informatics Forum building. The lab consisted of individual single rooms. Each room was equipped with an SAMSUNG 2043 screen monitor and a set of DT770 PRO headphones. Every participant would be arranged into one of the single rooms. The experiment was carried out by applying an online webpage. All the voices would come out from the headphones throughout the experiment, and the volume had been set into an adequate loudness to the listeners. No participants have complained about the sound volume.

3.3.1 Producing wav files

For intelligibility task and comprehension task, all wav files of SUS sentences and news passages had been produced by natural speech and the two synthesizers HTS-2008 and Multisyn. Since in intelligibility task, the wav files were generated by using every single sentence, the news passages used in the comprehension test were also synthesized into several single sentences for consistency. The sentences in the comprehension test were concatenated together into a passage afterwards, assigned with a silence interval of about 500 milliseconds between sentences.

There were some cases that needed to be carefully considered while producing synthesized speech, which the TTS systems could not identify the pronunciations 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 understood to listeners.

3.3.2 Pilot tests for comprehension task

Since the material used in the intelligibility test was the same as done in Blizzard Challenge, pilot tests for evaluating the intelligibility test were unnecessary. However, 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 difficulties of the text and questions, and text familiarity. 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 in the material, the wav files had been equally arranged into 6 different groups by using 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. In order to prevent the participants from having pressure on taking the exams, an announcement had been claimed beforehand indicating that they were not being tested but testing the systems.

The intelligibility task was taken first and then the comprehension task. This was done because more efforts were required while taking the comprehension task than intelligibility test, which participants needed to answer questions rather than type out what they heard. Therefore, it would be better for not depressing the listeners’ patience and willingness at 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 make the listening as few times as possible. For the comprehension task, listeners were only allowed to listen to each news article once, and then answered questions without note-taking technique. Also, two extra subjective questions were followed to each news article, asking about the participants’ confidence in completing the questions and their feelings of speech quality, scaling 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. [29] pointed out many researchers had found the participants would fail to sustain their attention after 20 to 35 minutes of doing the task. Due to the finding, participants were asked to have a 5-minute relaxing between the two tasks.

4. Results

4.1 Intelligibility task

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

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

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

In Pairwise Comparisons, as presented in Table 3, it reflects there are significant differences found between natural speech and HTS-2008 (p = 0.005), natural speech and Multisyn ($p <$ 0.001), and also 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 present that there are significant main effects when natural speech compares to HTS-2008, $F(1, 249)$ = 10.135, p = 0.002; and when HTS-2008 compares to 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 the HTS-2008 (M = 6.7%, SD = 11.4%), and the HTS-2008 is even better than 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) had been 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 are earned; if part of the answers are 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 could be possible. The examples of assessing the responses from open-ended questions had been 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 [17]. The reason for not taking a 2-point binomial scoring scale was because in real life comprehension, it was not always an all correct or wrong situation, as described by [30]. However, 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 point would be earned; reversely, if choosing the wrong answer, 0 points was graded. 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 neither in the interaction between systems and the question types. However, there is an obvious significant effect occurred in the question types, $F(1, 24) = 29.004$, $p < 0.001$. Therefore, the performance in open-ended questions is particularly 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. However, 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 to 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 between HTS-2008 and Multisyn system, *F*(1, 24) = 25.939, *p* < 0.001. Also, HTS-2008 performs a lot better (mean of error rate = 29.2%) than 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 to the displayed speech quality (Quality). Both of 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 satisfactory, 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 also 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 tests performance to 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 have occurred when the HTS-2008 compares to natural speech, *F*(1, 24) = 24.758, *p* < 0.001; and when Multisyn system compares to natural speech, *F*(1, 24) = 37.536, *p* < 0.001. While Quality compares to Confidence, two main effect is discovered in the interactions when the HTS-2008 compares to natural speech, *F*(1, 24) =

89.161, $p < 0.001$; when Multisyn compares with natural speech, $F(1, 24) = 73.059$, $p < 0.001$. Therefore, it can be concluded that the HTS-2008 is evaluated lower (M = 52.4%) than natural speech (M = 71.6%) in the subjective tests; and lower points is given to Multisyn (M = 52.2%) than to natural speech. Therefore, it is known that the natural speech has better results gained from the subjective tests, than the HTS-2008 and Multisyn systems.

In the Confidence test, it does not show any significant effect on the systems. This result indicates that listeners have equal confidence on natural speech, the HTS-2008, and Multisyn systems 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 to 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%), than the HTS-2008 (M = 48.8%) and Multisyn (M = 49.6%) systems. The results in Table 8 show no significance when HTS-2008 compares to Multisyn system. As a result of fact, in the subjective judgment of speech quality, natural speech is scored significantly higher than HTS-2008 and Multisyn systems. 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 the news articles are 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. However, the news articles produced by HTS-2008 and Multisyn systems are simply synthesized into individual sentences, without considering the context prosody factor. As stated by [31], listeners preferred the speech systems with higher prosody quality. Therefore, listeners have graded natural speech with the highest score, than HTS-2008 and Multisyn systems.

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 performances than Multisyn system. According to the initial hypothesis in this paper, predicting systems with higher achievement in the intelligibility task would also preserve better accomplishment in the comprehension task. In this case, we can estimate the three systems in the comprehension task might have the same rankings as presented in intelligibility task. However, in the overall comprehension task performances, no significant effects are noticed within the three systems, which signify natural speech, HTS-2008, and Multisyn all have relatively identical understandable quality for listeners. The outcomes in the comprehension task are against with the results in intelligibility task, and violate the hypothesis. Although, it seems that the comprehension task in this study has also failed to distinguish various speech systems, this is mainly because that the three systems have reached to 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 significance observed in the three systems, which meant that listeners have equivalent confidences in completing comprehension task produced by the systems. This implied that the three systems have given identical comprehension quality to the listeners. In addition, the techniques required for evaluating intelligibility and comprehension is 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 paying attention on every single word that can be heard. During the process of comprehending, 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. [14] had notified that with sufficient linguistic cues, it will 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 systems are significantly different from each other in the 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 Multisyn system, can be taken as an intelligibility threshold reference for achieving high comprehensibility in speech synthesizers.

5.1.2 The influences 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, rather than multiple-choice questions. 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 presents that natural speech, HTS-2008, and Multisyn have the same comprehensibility. However, the open-ended questions are designed to be factual questions, and that make the process of answering the questions to be similar to the way in completing the intelligibility task. Both the open-ended questions and intelligibility task involve listening to the speech first, and then focus on the key words they can capture or understand. The only difference between them is the load of memory will be larger in open-ended questions, than in intelligibility task. As seen into the results of open-ended questions, the consequences are a little diverse from the results in the intelligibility task. In the open-ended questions, the performances in natural speech are identical with the HTS-2008, but are better than the Multisyn system. Whereas, the intelligibility task presents that natural speech is better than the HTS-2008, and Multisyn. In addition, even in the overall subjective tests and quality test show that natural speech has better achievement than HTS-2008. This may contribute to the reason that there were not 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, in the overall subjective tests and quality test, they are both consistent with the results specifying that the performances in HTS-2008 and Multisyn system are the same. In general, the entire experiment in present study has found that natural speech has greater consequences and performances than HTS-2008 and Multisyn systems.

5.2 Listeners' feedback and some suggestions for future studies

5.2.1 Listeners' feedback

In the intelligibility task, most of the participants found it interesting. Since the materials were all semantically unpredictable sentences, that would make up a lot of unexpected funny sentences. Still, some of the participants specified that there were a few words they seldom heard and seen in their life, and might lead to some misspelling or make up the spelling pronunciations. This problem had been solved in this study, which we allowed typos and homonyms while calculating the WER in the intelligibility task. They had also indicated that sentences with poor speech quality, it would be hard 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 displaying duration of news articles is a bit long for them to remember the all the information. Besides, the listeners had notified that if the article was 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 topic they were interested in, and answered more correctly on the questions. Some participants suggested that there should be an option of "do not know the answer" added into 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 all in all.

5.2.2 Suggestions and modifications for future works.

According to the feedback received from the participants, there are some things that can be modified in the comprehension design to make the task better. Firstly, 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 need to be applied before carrying out the main experiment. Furthermore, since each news article is with different topics, there is no guarantee that the degree of text complexity and familiarity will still be the same between each article. The word "text complexity" used right here means the degree of comprehension effort that need 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 cease the individual problems and increase the results' objectivity in 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 the HTS-2008, and HTS-2008 is proved to be greater than the Multisyn system. However, the results in the comprehension task present that the natural speech, HTS-2008, and Multisyn systems are with equal quality for listeners to comprehend. The explanation has been given in section 5.1.1, discussing the issue may lead to the reason that all the three systems obtain high enough intelligibility quality to be used in comprehending the news passages. Although the outcomes in the intelligibility task show that there are significant differences investigated within the three systems, their intelligibility have reached to 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 manifest that listeners consider natural speech is 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 Multisyn systems.

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 [2]. 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 the comprehension, rather than the process of it.

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

7. References

[1] C. Stevens, *et al.*, "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,* vol. 19, pp. 129-146, 2005.

[2] D. B. Pisoni, *et al.*, "Perception of synthetic speech generated by rule," in *Proceedings of the IEEE*, 1985, pp. 1665-1676.

[3] H. A. Sydeserff, *et al.*, "Evaluation of speech synthesis techniques in a comprehension task," *Speech Communication,* vol. 11, pp. 189-194, 1992.

[4] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common dataset," in *Proceedings of Interspeech 2005*, Lisbon, Portugal, 2005.

[5] R. A. J. Clark, *et al.*, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication,* vol. 49, pp. 317-330, 2007.

[6] H. Zen, *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of ISCA SSW6,* Bonn, Germany, 2007.

[7] T. Yoshimura, *et al.*, "Simultanious modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of Eurospeech*, 1999, pp. 2347-2350.

[8] Z. Heiga and T. Tomoki, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Proceedings of Interspeech 2005*, Lisbon, Portugal, 2005, pp. 93-96.

[9] S.-Z. Yu and T. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Signal Processing Letters,* vol. 10, pp. 11-14, 2003.

[10] R. A. J. Clark, *et al.*, "Festival 2 - build your own general purpose unit selection speech synthesiser," in *Proceedings of 5th ISCA Speech Synthesis Workshop,*

Pittsburgh, USA, 2004, pp. 173-178.

[11] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the ICASSP 1996*, Atlanta, USA, 1996, pp. 373-376.

[12] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proceedings of the Eurospeech 1997*, 1997, pp. 601-604.

[13] L. C. W. Pols, *et al.*, "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, 1998.

[14] C. Benoît, *et al.*, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, pp. 381-392, 1996.

[15] G. A. Miller and S. D. Isard, "Some perceptual consequences of linguistic rules," *Journal of Verbal Learning and Verbal Behavior*, vol. 2, pp. 217-228, 1963.

[16] K. Yorkston, *et al.*, "Comoprehensibility of dysarthric speech: Implications for assessment and treatment planning," *American Journal of Speech-Language Pathology*, vol. 5, pp. 55-66, 1996.

[17] K. C. Hustad, "The relationship between listener comprehension and intelligibility scores for speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 562-573, 2008.

[18] P. A. Luce, "Comprehension of fluent synthetic speech produced by rule," Indiana University, Bloomington, IN 47405, Research on Speech Perception Progress Report 7, 1981.

[19] A. Salasoo, "Cognitive Processes and comprehension measures in silent and oral reading," Speech Research Laboratory, Indiana University, Bloomingtion, IN 47405, Research on Speech Perception Progress Report 8, 1982.

[20] D. B. Pisoni, *et al.*, "Perceptual evaluation of synthetic speech: Some considerations of the user/System interface," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, 1983, pp. 535-538.

[21] J. V. Ralston, *et al.*, "Comprehension of synthetic speech produced by rule," Speech Research Laboratory, Indiana University, Bloomington, IN47405, Research on Speech Perception Progress Report 15, 1989.

[22] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16, pp. 225-244, 1995.

[23] H. C. Nusbaum, *et al.*, "Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability," Speech Research Laboratory, Indiana University, Bloomington, IN47405, Research on Speech Perception Progress Report

10, 1984.

[24] H. Nusbaum, *et al.*, "Measuring the naturalness of synthetic speech," *International Journal of Speech Technology,* vol. 1, pp. 7-19, 1995.

[25] J. Terken and G. Lemeer, "Effects of segmental quality and intonation on quality judgments for texts and utterances," *Journal of Phonetics,* vol. 16, pp. 453-457, 1988.

[26] C. R. Paris, *et al.*, "Linguistic cues and memory for synthetic and natural speech," *Human Factors,* vol. 42, pp. 421-431, 2000.

[27] V. Karaiskos, *et al.*, "The Blizzard Challenge 2008," in *Proceedings of the Blizzard Challenge 2008 workshop* Brisbane, Australia, 2008.

[28] J. Lai, *et al.*, "The effect of task conditions on the comprehensibility of synthetic speech," presented at the CHI Letters, 2000.

[29] C. Delogu, *et al.*, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication,* vol. 24, pp. 153-168, 1998.

[30] K. C. Hustad and D. R. Beukelman, "Listener coomprehension of severely dysarthric speech: Effects of linguistic cues and stimulus cohesion," *Journal of Speech, Language, and Hearing Research,* vol. 45, pp. 545-558, 2002.

[31] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," *Language and Speech,* vol. 40, 1997.