Bulletin of the Psychonomic Society
1982, Vol. 20(4), 203-206

Recall of passages of synthetic speech

JAMES J. JENKINS and LYNNE D. FRANKLIN
University of Minnesota, Minneapolis, Minnesota 55455

Memory for synthetic speech versions of grade school-level materials was tested in two studies. In Experiment 1, two different versions of three simple stories were recorded in synthetic speech. The versions differed in prosody, one employing hand-applied pitch and stress and the other employing random stress. Free recall of the stories showed no consistent difference in performance as a function of the intonational pattern used. Experiment 2 compared the recall of a simple biographical sketch presented in either natural speech or synthetic speech. A sentence-by-sentence dictation test showed little difference in intelligibility of the texts, and the difference disappeared with minimal practice on synthetic speech. Free recall of the entire passage showed synthetic speech to be disadvantaged only in the case of nonpracticed listeners. Again, minimal practice with synthetic speech dispelled the differences.

Synthetic speech is one of the current vigorous applications of computer technology. Despite its potential for application, however, little is known concerning the comprehension and retention of continuous passages of synthetic speech.

Evaluations of synthetic speech would not only be useful and timely, but they might also shed light on the possible reasons for the well known acoustic redundancy of natural speech (e.g., Cherry, 1978; Miller, 1951). We may hypothesize that one reason for the redundancy in natural speech is that it eases the load on attentional resources, allowing the listener to focus on the meaning of what is being said rather than on the acoustic signal itself. Because synthetic speech usually lacks this acoustic redundancy (ordinarily being composed of some set of minimal cues), it is possible that primary attention must be given to phoneme identification, thereby limiting the amount of higher order processing (lexical, syntactic, and semantic) that might normally be employed in comprehending each sentence and relating it to the meaning of the total text.

Most research on synthetic speech has focused upon the intelligibility of individual words or words embedded in single sentences. For example, Nye and Gaitenby (1973) used a modified rhyme test to identify consonantal substitution errors in single words, finding an overall intelligibility error rate of 7.6% for synthetic speech and 2.7% for natural speech. In a second study, Nye and Gaitenby (1974) employed a dictation task with 200 syntactically normal but semantically anomalous sentences constructed from a pool of 252 frequently used words. Their results indicated that when subjects were asked to recall the four key words from each sentence directly after presentation, the overall error rate jumped to 22% for synthetic speech and rose only to 5% for natural speech. These data support the reasonable conclusion that synthetic materials in a nonredendant context are more difficult to recall than would be predicted from their observed intelligibility when they are presented as single words with a closed response set.

Schmidt-Nielsen (Note 1) studied comprehension of, and preference for, several versions of Votrax synthesis that differed in prosody. Although the various assignments of prosody differed greatly in rated preference, no reliable differences were found in the comprehension of the different versions of synthetic speech.

A more recent and extensive analysis of synthetic speech has been conducted by Pisoni and Hunnicutt (1980) in their evaluation of the MITalk text-to-speech system. Pisoni and Hunnicutt first evaluated MITalk in a fashion parallel to the Nye and Gaitenby (1973, 1974) studies described above. The modified rhyme test gave an error rate of 6.9% for synthetic speech and .6% for natural speech; a set of Harvard psychoacoustic sentences yielded 6.8% errors for synthetic speech and .8% errors for natural speech, and a set of the Nye and Gaitenby anomalous sentences showed 21.3% errors for synthetic speech and 2.3% errors for natural speech. It is obvious that these data, even though they are for a very different synthesis system, are remarkably parallel to the results of the Nye and Gaitenby studies for the Haskins Laboratories synthesis.

Pisoni and Hunnicutt (1980), like Schmidt-Nielsen (Note 1), also evaluated the comprehension of synthetic speech by synthesizing passages taken from reading

This research was supported by Grant MH-21153 from the National Institute of Mental Health to James J. Jenkins and Winifred Strange. The second author is a trainee of the Center for Research in Human Learning, supported by the National Institute of Child Health and Human Development (Grant HD-00098). Special thanks are due to Astrid Schmidt-Nielsen for providing tapes of synthetic speech and to Haskins Laboratories and David Isenberg for the synthetic material used in the second experiment. We gratefully acknowledge the help of Janet Fridgen, who independently scored all of the recall materials to ensure reliability. Requests for reprints should be sent to James J. Jenkins, Psychology Department, University of South Florida, Tampa, Florida 33620.

  0090-5054/82/100203-04$00.65/0

comprehension tests. No significant differences were found between natural and synthetic speech.

In subsequent studies Pisoni (1981) showed that synthetic speech tokens were more poorly classified on a lexical decision task than were natural speech tokens. Finally, Feustel, Luce, and Pisoni (1981) conducted a set of studies of short-term memory for word lists in natural and synthetic speech. When the investigators used a memory preload technique (loading short-term memory with zero, three, or six digits before presenting a list of 15 words), they found evidence suggesting that the synthetic word list interfered with the digit recall more than the natural word list did. The implication of these studies is that the activity of perceiving synthetic speech may put increased processing demands on short-term memory.

As a first step in evaluating the effectiveness of continuous synthetic speech, we undertook to study its effect on the recall of simple verbal materials. We reasoned that a direct and persuasive assessment of the efficiency of synthetic speech might be found in the listener's ability to recall the gist of sizable passages of verbal material. Recall is known to be influenced by apprehension of organization and structure and to be sensitive to semantic processing. In addition, recall tasks avoid the subject-matter prompting that is furnished by multiple-choice items.

EXPERIMENT 1

For the first experiment, we made use of some of the stories (mentioned above) that had been prepared by Dr. Astrid Schmidt-Nielsen at the Naval Research Laboratory in Washington, D.C. Dr. Schmidt-Nielsen was kind enough to lend us examples of the most preferred and the least preferred materials from her study for use in our research.

Method

Subjects. The subjects were 41 undergraduate students at the University of Minnesota who received course credit for their participation. Subjects were native speakers of English who had no prior experience listening to synthetic speech.

Materials. Stimulus materials were based on a Votrax synthesis (Votrax VS 6.0). The most preferred materials enhanced the phonetic synthesis with pitch, stress, and timing applied "by hand" to produce the most acceptable pattern of intonation for the texts. The least preferred materials were based on the same phonetic synthesis, except that stress was randomly applied to the syllables and timing differences were not introduced. It was quite obvious to even the untrained ear that the former synthesis was quite acceptable and the latter synthesis sounded highly artificial. (Of six versions evaluated by Schmidt-Nielsen, listeners ranked the most preferred version at an average of 1.35 and the least preferred version at an average of 4.65.)

For the most preferred material, stress was hand placed in agreement with correct English pronunciation. In addition, timing rules were employed. These included changes in syllable duration based on stress level and stress patterns within sentences. Specifically, vowel duration, known to be an important acoustic parameter in the perception of stress, was manipulated such that stressed vowels were assigned slightly longer durations than unstressed vowels.

Three stories from a reading comprehension test (at grade school level) were used in the study. Twenty-two subjects heard the least preferred versions, and 19 heard the most preferred versions. Listeners heard the stories in order from the easiest to the most difficult. After each story, listeners were asked to recall the material in the story. They were told to write down all the facts or ideas of the story and were assured that verbatim recall was not essential. All stories played to a given listener were of the same type; that is, they were all random stress versions or all hand-placed stress, pitch, and timing versions. The first story contained 27 scorable ideas, the second story contained 53 scorable ideas, and the last story contained 48 scorable ideas.

The tapes were played on a Revox tape recorder to small groups of subjects in a quiet room. Subjects were given as much time as they needed to write their recall of each story.

Results

The recall data are shown in Table 1. Means and standard deviations are given for the recall of each story under each condition. The percentage data posted beside the means perhaps provide the quickest comprehension of the results. It is apparent that the stories do proceed from easy to more difficult, as intended. It is also apparent that there is no clear superiority of the hand-applied pitch and stress over the random stress version of the stories. Significance tests confirm the absence of a clear outcome. No significant difference was found between the groups for Story 1 [$t(39) = .15$]. For Story 2, the random stress group actually performed better than the group listening to the version with hand-applied pitch and stress, although not significantly so [$t(39) = 1.1$]. For Story 3, the group with the more natural pitch and stress pattern was superior, although the results did not reach an acceptable statistical level [the 6% level, $t(39) = 2.0$].

These data suggest that a normal pattern of intonation is not critical in obtaining reasonable recall with simple synthetic speech materials. In recall, as in Schmidt-Nielsen's (Note 1) comprehension testing, great individual differences were found. Most surprising, however, was the fact that on one occasion (Story 2), the group that was thought to be handicapped by the presentation actually achieved a higher mean. Of course, it is possible to argue that both kinds of synthesis were not "equally good," but that they were simply "equally bad." That is, perhaps, all the recall scores are "on the floor." Unfortunately, we had not included a natural speech

Table 1
Means and Standard Deviations for the Recall of Simple Stories with Random Stress vs. Hand-Applied Stress

		Random Stress (N = 22)		Hand-Applied Stress (N = 19)	
Story	Possible	Mean	SD	Mean	SD
1	27	15.1 (56%)	4.0	15.0 (55%)	3.7
2	53	24.3 (46%)	8.6	21.7 (41%)	6.1
3	48	15.2 (32%)	7.8	19.3 (40%)	4.8

control in this study, so we have no way of assessing this.

The next experiment was designed to compare recall of synthetic speech with that of natural speech and, in addition, to make the comparison slightly more realistic by giving one group a small amount of practice in listening to synthetic speech prior to the experiment itself.

EXPERIMENT 2

The second experiment was directed toward a comparison of two kinds of materials, natural speech and synthetic speech, and two conditions of listening to synthetic speech, with practice and without practice. Because this study involved a new synthesis, the materials were tested for word-for-word intelligibility as the words occurred in single sentences. Obviously, it would be of little interest to show that the synthesis was difficult to recall if the listeners had not understood the message in the first place.

Method

Subjects. The subjects were 84 undergraduate students at the University of Minnesota who received course credit for participation. All subjects were native English speakers with no prior exposure to synthetic speech.

Materials. The materials were prepared by Dr. David Isenberg and J. J. Jenkins at Haskins Laboratories, using the FÖVE synthesis-by-rule system. This system accepts phonetic text as an input and drives an ÖVE speech synthesizer via phonetic-to-acoustic rules developed by Ingemann (1978, 1979). Although the phonetic-acoustic synthesis is almost completely rule governed, the operator must make choices in the system with respect to intonation. Accordingly, intonation was set by hand, "artistically." The material was a biography taken from the children's section of a New Haven newspaper.

A natural speech version of the biography was prepared after the synthetic version was complete. Dr. Isenberg listened to each synthetic sentence on earphones and repeated it with comparable timing and intonation pattern to produce a closely matched version of the synthetic story in natural speech.

Intelligibility testing. Intelligibility was tested by a dictation task. The text was played sentence by sentence, with long pauses between the sentences to allow subjects to write the sentences as accurately as they could. Three groups of 14 subjects each were tested. One group heard the natural speech version of the sentences; two groups heard the synthetic text version of the sentences. One of the latter groups was naive to synthetic speech. One week before being tested on the experimental materials, the other group had practiced on 20 sentences of synthetic speech (on an unrelated topic), produced by a comparable rule system. All of the subjects took the dictation test twice; the second trial immediately followed the first.

Results

Table 2 shows the results of the word-for-word scoring of the dictation test. There were a total of 169 words. From the percentage data, it is apparent that all of the texts were highly intelligible but that natural speech was slightly more intelligible than synthetic speech. Trial 1 results showed significant differences between the natural speech group and the synthetic group without practice [$t(26) = 4.0$, $p < .001$], as well as between the natural speech group and the synthetic group with practice [$t(26) = 3.4$, $p < .01$]. By the second trial, however, the differences were exceedingly small; the only significant difference appeared between the natural speech group and the synthetic group without practice [$t(26) = 2.2$, $p < .05$].

Table 2
Means and Standard Deviations for the Pam Dawber Text Dictation Condition, Word-for-Word Scoring

	Natural (N = 14)		Synthetic (N = 14)		Practice (N = 14)	
Trial	Mean	SD	Mean	SD	Mean	SD
1	161.9 (96%)	5.5	149.5 (89%)	9.6	152.4 (90%)	8.3
2	166.1 (98%)	2.4	161.2 (95%)	7.7	162.8 (96%)	4.4

Note–169 possible. Trial 1: N > S, P. Trial 2: N > S.

Table 3
Means and Standard Deviations for the Pam Dawber Text Dictation Condition, Idea Scoring

	Natural (N = 14)		Synthetic (N = 14)		Practice (N = 14)	
Trial	Mean	SD	Mean	SD	Mean	SD
1	35.7 (97%)	1.7	33.5 (91%)	1.7	35.0 (95%)	1.5
2	36.9 (99%)	<.1	35.6 (96%)	1.3	36.0 (97%)	1.3

Note–37 possible. Trial 1: N > S; P > S.

Because the recall part of the experiment was to be scored by ideas, the dictation test was also scored in that fashion. Table 3 shows the results on the dictation task when it was scored on the basis of the 37 possible "idea units."

On Trial 1, the synthetic group without practice recorded fewer ideas than either the natural speech group [$t(26) = 3.3$, $p < .01$] or the synthetic group with practice [$t(26) = 2.4$, $p < .05$]. On the second trial, however, there were significant differences between the natural speech group and both the synthetic group with practice [$t(26) = 2.5$, $p < .05$] and the synthetic group without practice [$t(26) = 3.4$, $p < .01$]. Due to the miniscule variance of the natural speech group, however, these statistical tests of second-trial results must be regarded with suspicion. Obviously, the natural speech group was nearly at ceiling. In order to derive a more general error term, a repeated-measures ANOVA was carried out on these data. It indicated a main effect for groups [$F(2,39) = 7.1$, $p < .01$] and trials [$F(1,39) = 44.2$, $p < .001$]. The interaction was not significant. The results of orthogonal contrasts on the means between the groups for Trial 1 (using the pooled variance estimate) indicated a significant difference between the natural speech group and the synthetic group without practice [$t(39) = 3.4$, $p < .01$] and between the two synthetic speech groups [$t(39) = 2.2$, $p < .05$]. Data for Trial 2 indicated no significant differences between groups. Collapsing over trials, only the natural speech group and the synthetic group without practice were found to differ [$t(39) = 3.8$, $p < .001$].

These data suggest that the synthesis is reasonably intelligible and that almost all of the basic ideas are

Table 4
Means and Standard Deviations for the Pam Dawber Text Free Recall Condition, Idea Scoring

	Natural (N = 14)		Synthetic (N = 14)		Practice (N = 14)	
Trial	Mean	SD	Mean	SD	Mean	SD
1	17.1 (46%)	4.4	11.0 (30%)	4.6	16.1 (44%)	3.0
1&2	27.6 (75%)	3.0	21.7 (59%)	7.4	25.5 (69%)	2.8

Note–37 possible. Trial 1: N > S; P > S. Trials 1 and 2: N > S.

available to the listeners. Furthermore, if the listener has a little practice with synthetic speech before the experiment, the synthetic text is virtually as intelligible as the natural speech text.

The free recall task employed three groups of 14 listeners each, paralleling the dictation study. One group heard the natural speech version of the text. Two groups heard the synthetic version of the text. One of these groups (practice) had been through a recall session with 20 sentences of unrelated synthetic text 1 week earlier. The other group was naive to synthetic speech. Subjects heard the tape recording appropriate to their group as a continuous text. At the end of the recording, the listeners were asked to recall as much of the text as they could. When they had finished their recall, the tape was played again. At the end of the text, the listeners were asked to recall any additional facts they had missed in their first attempt (which was still available to them).

Table 4 gives the means and standard deviations for the three groups when their recall was scored in terms of idea units. The statistical tests disclose no reliable difference between the natural speech group and the synthetic group with practice either on the sum of the two trials or on the first trial separately. The group hearing synthetic speech for the first time performed significantly more poorly on the first trial than either the natural speech group [$t(26) = 3.5$, $p < .01$] or the synthetic group with practice [$t(26) = 3.5$, $p < .01$]. The synthetic group without practice also performed significantly more poorly than the natural speech group on the two trials combined [$t(26) = 2.7$, $p < .05$].

DISCUSSION

The results of the present set of studies indicate that synthetic speech is remarkably robust. The first experiment showed that unnatural-sounding synthetic speech can yield relatively good recall of simple text materials compared to more natural-sounding synthetic speech. The second experiment found that with even a little practice, synthetic speech seems to be as intelligible and as memorable as natural speech.

It is important to remember that the materials used in these experiments were very simple texts, far below the intellectual level of the listeners. This fact may have hidden more persistent and real differences between natural and synthetic speech texts. The next step is to test the generality of the present findings with more complicated materials and in situations that increase task demands on the listeners. More realistic tests are now called for. At the same time that we move toward this further research, however, we must pause for a moment to admire the accomplishments of speech synthesis to date. Even with simple materials, it is not a trivial feat to achieve the levels of intelligibility and memorability displayed in the experiments presented here.

REFERENCE NOTE

1. Schmidt-Nielsen, A. *Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program* (Report 8015). Washington, D.C: Naval Research Laboratory, September 9, 1976.

REFERENCES

CHERRY, C. *On human communication* (3rd ed.). Cambridge, Mass: M.I.T. Press, 1978.

FEUSTEL, T. C., LUCE, P. A., & PISONI, D. B. Capacity demands in the short-term memory for synthetic and natural word lists. *Journal of the Acoustical Society of America,* 1981, **70**, S98. (Abstract)

INGEMANN, F. Speech synthesis by rule using the FÖVE program. *Haskins Laboratories Status Report on Speech Research,* 1978, **SR-54**, 165-173.

INGEMANN, F. The contributions of natural durations to speech synthesized by FOVE rules. *Haskins Laboratories Status Report on Speech Research,* 1979, **SR-58**, 177-184.

MILLER, G. A. *Language and communication.* New York: McGraw-Hill, 1951.

NYE, P. W., & GAITENBY, J. H. Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). *Haskins Laboratories Status Report on Speech Research,* 1973, **SR-33**, 77-91.

NYE, P. W., & GAITENBY, J. H. The intelligibility of synthetic monosyllable words in short syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research,* 1974, **SR-37/38**, 169-190.

PISONI, D. B. Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America,* 1981, **70**, S98. (Abstract)

PISONI, D. B., & HUNNICUTT, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In *1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing,* April 1980.

(Received for publication July 9, 1982.)