Published in final edited form as:
Comput Speech Lang. 1987 ; 2(3-4): 303–320.

Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility

David B. Pisoni, **Laura M. Manous**, and **Michael J. Dedina**
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN 47405, U.S.A

Abstract

A sentence verification task (SVT) was used to study the effects of sentence predictability on comprehension of natural speech and synthetic speech that was controlled for intelligibility. Sentences generated using synthetic speech were matched on intelligibility with natural speech using results obtained from a separate sentence transcription task. In the main experiment, the sentence verification task included both true and false sentences that varied in predictability. Results showed differences in verification speed between natural and synthetic sentences, despite the fact that these materials were equated for intelligibility. This finding suggests that the differences in perception and comprehension between natural and synthetic speech go beyond segmental intelligibility as measured by transcription accuracy. The observed differences in response times appear to be related to the cognitive processes involved in understanding and verifying the truth value of short sentences. Reliable effects of predictability on error rates and response latencies were also observed. High-predictability sentences displayed lower error rates and faster response times than low-predictability sentences. However, predictability did not have differential effects on the processing of synthetic speech as expected. The results demonstrate the need to develop new measures of sentence comprehension that can be used to study speech communication at processing levels above and beyond those indexed through transcription tasks, or forced-choice intelligibility tests such as the Modified Rhyme Test (MRT) or the Diagnostic Rhyme Test (DRT).

Over the past six years, numerous studies on the perception of synthetic speech have been conducted in our laboratory at Indiana University (Pisoni, 1982; Nusbaum & Pisoni, 1985; Pisoni, Nusbaum & Greene, 1985). The bulk of these studies have focused on measures of segmental intelligibility, such as identification of isolated words and recognition of words in sentences (e.g., Egan, 1948; House, Williams, Hecker & Kryter, 1965; Nye & Gaitenby, 1973). Results from these studies of phoneme and word perception have shown that synthetic speech is consistently less intelligible than natural speech (Greene, Logan & Pisoni, 1986). This finding was observed for a variety of synthesis systems ranging from very low-quality, low-intelligibility products, such as the ECHO and Votrax Type N Talk, to extremely natural sounding speech with very high intelligibility such as DECtalk and the Prose 2000.

Most perceptual studies dealing with segmental intelligibility have not addressed the issue of comprehension processes involved in understanding the linguistic content of the message. In tests of segmental intelligibility, such as the ones we have carried out, subjects are not

Correspondence to: David B. Pisoni.

required to extract or compute the meanings of utterances in order to make appropriate responses. They can carry out the task based on their discrimination of the acoustic–phonetic properties of the speech alone, without fully understanding what they are listening to, or making a response that is based on the comprehension of the message. Depending on the type of comprehension test employed, subjects must use other information to generate a correct response above and beyond the acoustic–phonetic cues in the speech waveform.

To date, relatively little work has been done to examine how listeners comprehend synthetic speech produced automatically by text-to-speech systems. Speech quality as well as overall segmental intelligibility of the speech signal are certainly important factors involved in spoken language comprehension. Yet, additional consideration must also be given to the contribution of higher sources of knowledge in “understanding” the message and responding appropriately to the truth-value of sentences.

The few studies that have been conducted to examine comprehension of natural and synthetic speech have produced equivocal results, making it difficult to draw any general conclusions about the comprehension process. In one early study, McHugh (1976) assessed comprehension of synthetic and natural speech using passages selected from a standardized reading comprehension test. Prosodic information was manipulated by presenting six different stress variations of the synthetic speech from the Votrax synthesizer, along with a natural speech control condition. Subjects’ performances showed no significant differences across the seven conditions. McHugh concluded that the test she used was, therefore, too sensitive to individual differences in performance to reveal any differences between the various experimental conditions that were tested.

In studies carried out in our laboratory, Pisoni (1987) and Pisoni & Hunnicutt (1980) studied the comprehension of natural speech and synthetic speech produced by MITalk, a text-to-speech system developed at MIT (Allen, 1981; Allen, Hunnicutt & Klatt, 1987). Listening comprehension was compared to reading comprehension for identical passages, using multiple-choice questions taken from standardized reading comprehension tests. The results demonstrated that naïve listeners were able to comprehend passages of synthetic speech at levels comparable to subjects who had either heard passages of natural speech or who had read the passages and answered the same questions after presentation of each passage.

In another study, using passages of connected speech, Jenkins & Franklin (1981) examined comprehension of natural speech and synthetic speech produced by a Votrax text-to-speech system, using a free recall task and a sentence dictation procedure. One group of subjects transcribed a passage presented one sentence at a time. Another group of subjects listened to the entire passage and then attempted to recall the information just presented, in a free recall format. Once again, the results showed little difference in performance between natural and synthetic speech. Apparently, the behavioral measures used to assess comprehension in these studies were too gross and too insensitive to reveal differences between various types of speech.

More recently, Schwab, Nusbaum & Pisoni (1985) included listening comprehension passages and true–false questions, along with several other tests, to study the effects of perceptual learning on the perception of synthetic speech. As in the previous studies, the particular comprehension task we used did not reveal any effects of training or any differences between natural and synthetic speech. These results were surprising because all of the other tests (e.g., identification of isolated words, recognition of words in fluent sentences) used to assess performance in this study showed highly significant effects of training on the perception of synthetic speech.

Following up on these earlier comprehension studies, Moody & Joost (1986) have recently examined listener comprehension rates for synthesized speech using DECtalk, digitized speech using 9·6 and 2·4 kbps LPC algorithms, and natural recorded speech. Passages and multiple-choice questions were selected from standardized verbal exams such as the SAT and GRE. Their results showed significant differences in question–answer performance for synthetic speech and 2·4 kbps LPC digitized speech compared to natural speech. The difficulty of the passage affected comprehension rates for all passages, regardless of the type of speech signal used. However, Moody & Joost observed an unusual interaction between passage difficulty and speech type in their study. When subjects listened to more difficult information in some passages, differences in performance between the natural speech group and the synthetic speech group were not observed. However, when the comprehension materials were easy, significant differences between the natural and synthetic speech groups emerged.

It is not immediately obvious to us how one would account for these findings, given the results of an earlier study by Luce, Feustel & Pisoni (1983), which showed increased error rates in serial recall when capacity demands of the task were increased. We do not know of any current theory of human information processing or language comprehension that would predict the results observed by Moody & Joost. If there is some relationship between comprehension difficulty and signal quality, then differences in performance among natural speech, synthetic speech, and digitally vocoded speech should emerge more robustly under experimental conditions in which there are greater capacity demands on the processes used in perception or comprehension. Resolution of this problem obviously awaits additional research on comprehension, using long passages of connected speech that have been specifically designed to differ in comprehension difficulty. For the present, we simply wish to point out that research on comprehension of synthetic speech continues to yield equivocal results that are difficult to integrate with other findings reported in the literature. When a situation like this arises, it is often useful to examine some of the commonalities and differences in the experimental procedures that have been used in this research, and to consider alternative techniques that may be used to approach the same general problem in different ways.

When considered together, all of the previous studies on comprehension of synthetic speech have a number of similarities. Firstly, they all used post-perceptual measures to index differences in comprehension. It is well known in the comprehension literature that post-perceptual measures are affected by a variety of subject strategies that rely on numerous sources of knowledge, in addition to the linguistic information contained in the input signal. Secondly, all of these studies used multiple-choice or true–false question/answering tasks or recall measures which often encourage subjects to exploit their real-world knowledge to solve the task. And, finally, they used accuracy measures to index processing load instead of response latencies. The consistent failure to find differences in perception between natural speech and several kinds of synthetic speech using these measures suggests the need for much more sensitive methods of measuring ongoing processing activities. One such method is the sentence verification task, which has been used extensively in previous psycholinguistic investigations of the language comprehension process. In these procedures, subjects are required to judge whether a sentence is “true” or “false”. Response latency is used as the primary dependent variable.

Sentence verification has been used for many years to assess processing activities in studies on language perception and comprehension (for a review, see Clark & Clark, 1977). In one of the earliest studies using this procedure, Gough (1965, 1966) found that sentence verification time varied as a function of grammatical form. Reaction times were shorter for active as opposed to passive sentences, affirmative as opposed to negative sentences, and

true as opposed to false sentences. Collins & Quillian (1969, 1970) and Conrad (1972) have used sentence verification to study the organization and retrieval of semantic knowledge about words in long-term memory (see Chang, 1986, for a recent review). Both studies used response time as a measure to infer the level of processing required to verify information contained in various types of sentences, such as “a canary is a bird” or “a canary has wings”. More recently, Larkey & Danly (1983) used sentence verification to investigate the role of prosody in comprehension of digitally vocoded natural speech. They found that subjects were, on average, 48 ms slower in responding to sentences with a monotone pitch than to sentences with the original prosodic contour left intact.

In a recent study carried out in our laboratory, Manous, Pisoni, Dedina & Nusbaum (1985) used the sentence verification task to investigate differences in comprehension between natural speech and synthetic speech generated by five different text-to-speech systems. They found that response latencies to verify short sentences increased as segmental intelligibility of the speech decreased. Specifically, the results yielded a reliable rank-ordering of the different synthetic voices in which level of performance corresponded to the quality of segmental information for each type of speech. That is, performance on the sentence verification task for the voices tested, followed the same pattern observed in earlier standardized tests of segmental intelligibility (Greene, Logan & Pisoni, 1986). These findings suggest that the early stages of the comprehension process depend primarily, if not exclusively, on segmental intelligibility. However, it is possible that other processes are also affected by the quality of the initial acoustic–phonetic input in the speech signal. Differences in the early stages of perceptual analysis of the input may cascade up the processing system and impact on other processes more closely related to comprehension (see also, Streeter & Nigro, 1979).

The present study was designed to examine this issue more closely and to dissociate effects due to segmental intelligibility from those related to comprehension processes. By controlling the level of intelligibility of the speech, we hoped to assess the comprehension process more directly and to draw inferences about processing activities that are not confounded with initial differences in segmental intelligibility. To accomplish this we matched high-quality synthetic speech produced by DECtalk with natural speech in terms of segmental intelligibility. We then used the sentence verification task to compare performance for these two types of stimulus materials, using test sentences that varied in length and semantic predictability. If the differences in perception between natural speech and very high-quality synthetic speech are not due only to segmental intelligibility, then we would expect to find differences in response times in a verification task, even though the error rates and transcription scores for the two types of speech were comparable. Such a finding would be a significant demonstration that the perception and comprehension of synthetic speech differs in important ways from the processing of natural speech (Pisoni, 1982). Moreover, such a finding with stimulus materials controlled for segmental intelligibility would suggest that cognitive processes, related to comprehension and understanding of the message, are also affected by the initial quality of the acoustic–phonetic input in the speech signal.

If synthetic speech is indeed more difficult to comprehend, in some general sense, than natural speech, then this difference should be influenced by other factors that affect speech perception and spoken language comprehension. In order to investigate this hypothesis we manipulated the predictability of the last word in the test sentences. In low-predictability sentences, less contextual information is available from earlier context to facilitate the perceptual process. In this case, listeners must rely more heavily on the acoustic–phonetic input in these sentences; thereby drawing scarce processing resources away from high-level comprehension processes. Assuming that the human speech processing system has only

limited processing capacity at its disposal, we expect that, if synthetic speech is more difficult to understand than natural speech, a manipulation of predictability would have a larger effect on synthetic rather than natural speech. In addition to manipulating sentence predictability, we also varied sentence length as a rough index of syntactic complexity. We expected to find interactions of these two variables with the voice manipulation. Using sentence length as an index of syntactic complexity, we expected to find that long sentences would be more difficult to process than short ones and that this effect would be reliably greater for synthetic rather than natural speech.

1. Method

1.1. Subjects

Subjects were either volunteers who were paid $3.50 for their participation in this study or introductory psychology students who participated to fulfill a course requirement. Subjects were drawn from the same general population at Indiana University. An equal number of subjects from these two groups participated in each condition of the experiment. All were native speakers of English and reported no history of a speech or hearing disorder at the time of testing. None of the subjects had any extensive experience in listening to synthetic speech before the present experiment.

1.2. Stimuli

In the first phase of the experiment, test items were specifically developed to vary in their semantic predictability. These materials were generated by having subjects provide the final word to complete 100 three-word and 100 six-word sentence frames. Examples of these stimuli are given in Table I. For half of the sentences of each length, subjects were instructed to create true sentences, for the other half they were required to construct false sentences; 40 subjects participated in this phase of the experiment.

The data from this task were scored in terms of response frequency for each item. The sentences were then categorized by response predictability. Sentences for which a high frequency of subjects gave the same response and for which there was only one response of high frequency were labelled “High-Predictability”. Out of 40 subjects, 25 or more had to respond with the same word in order for a sentence to be classified as “High-Predictability”. “Low-Predictability” sentences were defined as sentences for which there was a unique response, that is, sentences for which only one subject gave that particular response. Examples of “High-” and “Low-Predictability” sentences are shown in Table II. The entire set of stimulus materials is given in the Appendix.

The second phase of the experiment involved intelligibility testing of the sentences that were generated using synthetic speech produced automatically by rule. This phase was designed to match synthetically produced test sentences with natural test sentences for segmental intelligibility. The sets of “High-” and “Low-Predictability” sentences obtained in Phase 1 were recorded on audio tape using the DECtalk Version 2·0 text-to-speech system. These sentences were then presented to subjects in a transcription task.

Twenty-four additional subjects listened to the sentences and transcribed each one as accurately as possible with paper and pencil. Transcriptions were scored for exact phonemic match to the original sentences. Spelling errors were ignored unless they affected the meaning of the sentence (e.g., “medal” for “metal”).

Based on the data obtained in Phase 2, 40 true and 40 false sentences were selected for use in the main sentence verification task. For 77 of these sentences there were no transcription errors; each of the remaining three sentences had only one transcription error. Half of the

sentences selected were “High-Predictability” sentences; the other half were “Low-Predictability” sentences. In addition, half of the sentences of each type were three-word sentences and half were six-word sentences.

Additional tokens of each of the 80 test sentences were produced by a male talker (PAL). Both groups of test sentences, the synthetic speech and the natural speech materials, were low-pass filtered at 4·8 kHz, then digitized at 10 kHz using a 12-bit A/D converter and edited into individual stimulus files using a digital waveform editing program on a PDP-11/34 mini computer.

1.3. Procedure

Sixty subjects participated in the final phase of the experiment. Two to five subjects were run at a time in small groups. Each subject sat at a booth equipped with high-quality matched and calibrated headphones (Telephonics TDH-39) and a two-button response box. Stimulus presentation and response collection were controlled on-line by a PDP-11/ 34 computer. At the beginning of each session, the experimenter read the instructions aloud while the subjects simultaneously read a printed version in front of them. Subjects were told that they would hear one sentence on each trial and that their task was to determine if the sentence was “true” or “false”. Each group of subjects heard only one type of speech; half of the subjects listened to natural speech and half listened to synthetic speech. Sentence length and sentence predictability were within-subject factors.

Subjects were given four practice trials to familiarize them with the task and with the sound quality of the voice used in that particular condition. Following the practice trials, 80 experimental trials were presented. Test sentences were presented to subjects over headphones, via a 12-bit D/A converter. On each trial, subjects first heard a sentence and then made a forced-choice true/false response by pressing one of two appropriately labelled buttons on a response box. Subjects were instructed to respond as quickly and as accurately as possible when making their true/false decisions. After entering their response, subjects were also required to transcribe each sentence on a separate answer sheet using paper and pencil. This additional task was included to ensure that subjects had correctly encoded the test sentences after hearing them.

During the course of the experiment, the experimenter remained in the room to ensure that subjects were responding appropriately. Test trials were paced to the slowest subject in each group. Response latencies were measured using computer-controlled routines from the onset of each sentence to the subject’s response. The duration of each sentence which was available from previous measurements was then subtracted from the observed response latency to provide a measure of response time that was not contaminated by differences in stimulus length.

2. Results

A few of our subjects seemed to ignore our request that they respond quickly in the experiment. In order to reduce the subsequent variability in our data, we omitted from our final analyses the subjects whose average response latencies were greater than two standard deviations from the mean. Using this criterion, three subjects were eliminated, two from the natural speech group, and one from the synthetic speech group. We also eliminated one additional subject from the synthetic speech group, so that the same number of subjects were omitted from each experimental condition. This last subject had the slowest mean response time of the remaining subjects in the synthetic speech group. Thus, the final analyses reported below were based on data collected from 56 subjects.

2.1. Sentence transcription scores

In order to confirm that we had, in fact, successfully controlled for the segmental intelligibility of the sentences across the two sets of stimulus materials used in the experiment, we analyzed the effect of voice (natural *vs.* synthetic) on transcription accuracy. An analysis of variance on the transcription scores revealed no significant effect of voice for true sentences [$F(1,54) = 0{\cdot}96$, N.S.] or for false sentences [$F(1,54) = 0{\cdot}70$, N.S.]. The data were then analyzed using the two other dependent measures: (1) sentence verification accuracy, and (2) response latency. Separate analyses were carried out for each dependent measure to assess the effects of the three experimental manipulations—voice, sentence length, and sentence predictability. In carrying out these analyses, true and false sentences were always analyzed separately. The experimental design included three main effects: voice was a "between-subjects" factor, whereas sentence length and sentence predictability were "within-subjects" factors. Unless otherwise noted, the significance levels reported below are for the $p < 0{\cdot}01$ level of confidence.

2.2. Sentence verification accuracy

Figure 1 shows the verification error rates for true and false sentences. Overall, the error rates were quite low, demonstrating that subjects had little difficulty in understanding the sentences and carrying out the verification task with both natural and synthetic speech. Inspection of the error rates shown in Fig. 1 reveals several consistent effects of the experimental variables. For the true responses—displayed in the top panel of this figure—the observed error rates were consistently higher for low-predictability sentences than for high-predictability sentences. This was found for both natural and synthetic speech and was observed at each of the two sentence lengths used in the study. Analysis of variance confirmed these observations for the effects of predictability on true sentences [$F(1,54) = 38{\cdot}01$, $p < 0{\cdot}001$]. All other effects in analyses of the error rates for both true and false responses failed to reach significance.

Although there was a trend for the error rates to be slightly higher for synthetic speech, the differences were not reliable in either analysis of the true or false sentences. This result is not surprising, considering the procedures that were used to match sentences on intelligibility before the main verification experiment was carried out. The absence of an effect of voice in the analysis of the verification error rates is also consistent with the analyses of the transcription data described earlier, in which no differences were found in immediate recall between the natural and synthetic sentences. Thus, taken together, both sets of data—the transcription scores and the sentence verification error rates, suggest that subjects correctly encoded the stimulus materials at the time of input and that they successfully comprehended the linguistic content and meaning of the sentences. Although a reliable effect of sentence predictability was observed for the true sentences, the absence of a main effect for voice combined with the absence of any interactions with the voice manipulation suggests that the differences in the perceptual encoding between the natural and synthetic stimuli were minimal at best. In short, the expected outcome for both of these measures was observed.

2.3. Verification response latency

Response latencies were analyzed only for sentences that had been *both* verified correctly and transcribed correctly. Figure 2 shows the mean response latencies for true and false sentences in each of the conditions of the design.

Inspection of both panels in Fig. 2 shows several consistent effects for true and false response latencies. Firstly, as expected, there was a very prominent effect of sentence predictability on response latency. This is displayed in both panels of the figure. High-

predictability sentences were responded to much more rapidly than low-predictability sentences, and this effect was present for both true and false sentences, respectively. Table III shows the mean latencies for each of the four cells (collapsing across voice) in both conditions of the experimental design. Separate analyses of variance on the true and false responses revealed highly significant effects for sentence predictability, $F(1,54) = 121{\cdot}64$, $p < 0{\cdot}001$ and $F(1,54) = 32{\cdot}74$, $p < 0{\cdot}001$, respectively. No interactions were observed in either analysis for sentence predictability.

Secondly, a consistent effect of voice (i.e., natural *vs.* synthetic speech) can be observed in both panels of Fig. 2. Natural speech was consistently responded to more rapidly than synthetic speech. Table IV provides the mean latencies for the four cells (collapsing across predictability) in the experimental design for the true and false responses.

Although Fig. 2 displays this effect for both true and false responses, separate analyses of variance established that the effect of voice was only significant for the true responses [$F(1,54) = 9{\cdot}07$, $p < 0{\cdot}004$]. The ANOVA for the false responses produced a result that was in the expected direction from the data trends shown in Fig. 2, but it did not quite reach statistical significance [$F(1,54) = 3{\cdot}94$, $p < 0{\cdot}053$].

None of the other main effects or interactions reached statistical significance in either analysis of true or false responses and no interactions were observed with either of the two main variables (i.e., predictability and voice) that did reach significance. Thus, consistent and reliable differences in verification latencies between natural and synthetic speech were observed, even when the sentences were controlled for intelligibility. These results provide evidence against the claim that the observed differences were due to differences in segmental intelligibility between natural and synthetic speech or differences in encoding strategies at the time of input. The 200 ms overall mean difference in the response latencies between natural and synthetic speech found for the true responses suggests that some aspect of the comprehension process, other than perceptual encoding, is affected by the quality of the acoustic–phonetic input in the speech signal. It is clear from these findings that highly intelligible synthetic speech produced by rule still produces a decrement in performance, even when elaborate steps have been taken to experimentally control for differences in the initial level of intelligibility of the stimulus materials. The nature of these differences in the comprehension process will be considered below.

3. General discussion

The present investigation was designed to study the comprehension process using much more sensitive response measures than have been employed in earlier studies dealing with the perception and comprehension of synthetic speech produced automatically by rule. Using short meaningful three- and six-word sentences that were equated for segmental intelligibility, we found that response latencies in a sentence verification task were reliably faster for sentences produced using natural speech than the same sentences produced using high-quality synthetic speech generated by DECtalk. Thus, to a first approximation, we were reasonably successful in finding a comprehension task that would reveal meaningful differences in performance between natural speech and very high-quality synthetic speech. In the sections below we offer an account of these findings in terms of earlier work using the sentence verification task to study language comprehension processes.

As we noted in the introduction to this report, previous studies on the comprehension of synthetic speech have consistently failed to find reliable differences in performance between natural speech and several kinds of low-quality synthetic speech. Such findings have appeared anomalous to us because other measures of phoneme perception, word recognition and sentence transcription have all reliably discriminated, not only between natural and

synthetic speech, but more importantly between different kinds of synthetic speech, ranging from high-quality systems such as DECtalk to very poor-quality systems such as ECHO and Votrax (Greene, Logan & Pisoni, 1986). We discussed a number of criticisms of the specific experimental procedures used in these earlier studies, including some of our own research, and we suggested several alternatives to pursue in future work on this problem. The present experiment, which used a sentence verification task to study comprehension, was specifically designed with these criticisms in mind.

In addition to finding differences in the verification latencies between natural and synthetic speech, we also observed a reliable effect of sentence predictability on response latencies. This effect was found for both true and false responses and was extremely robust for natural and synthetic speech. As expected, high-predictability sentences were consistently responded to more rapidly than low-predictability sentences. To our surprise, however, we failed to find any reliable effects of sentence length on verification latencies. We also failed to find any interactions among the three experimental variables manipulated in this experiment. Contrary to our original expectations, we did not observe the predicted interactions between voice and sentence predictability on the one hand, and between voice and sentence length on the other. Both of these interactions would have demonstrated differential effects of these two structural variables on the comprehension of synthetically produced sentences. Precisely why we failed to find these results is unclear at this time; several suggestions will be considered below. Additional experimental manipulations will be needed to determine the locus of the observed differences between natural and synthetic speech in the language processing system. For the present time, however, the absence of the predicted interactions between voice and sentence predictability, and voice and sentence length, are important findings that merit further attention.

The present results differ from the earlier study by Manous *et al.* (1985) in a number of respects that are important to consider at this point. In the original SVT study, we found that sentence verification error rates and response latencies were strongly related to the segmental intelligibility of the particular text-to-speech system under study. However, differences in segmental intelligibility among the text-to-speech systems varied quite widely, and therefore the observed differences in the verification test could be attributed to a variety of factors, among which might be *real* differences in the comprehension process itself or simply differences in the initial levels of intelligibility of the systems. As it stands, our earlier study could not discriminate between the potential source, or sources, of the observed differences in either the verification error rates or the response latencies. Differences in the verification error rates suggest, however, that subjects probably did have some difficulty encoding many of the sentences, particularly those produced by the low-quality text-to-speech systems. Our analyses of the transcription data, collected after each sentence was verified, further suggested that this was a reasonable account of the differences. Thus, the most parsimonious explanation of the results of our earlier study was that the observed differences were probably due to difficulties at the time of encoding, because the initial level of intelligibility of the systems varied so widely (see Greene, Logan & Pisoni, 1986, for a summary of the intelligibility data for these systems).

With regard to the outcome of the present study, such an explanation would be difficult to maintain because the segmental intelligibility of the test sentences was very carefully controlled before the experimental data were collected. Moreover, the observed error rates for the verification responses were now extremely low, compared to the earlier study, and no reliable differences could be observed in the pattern of the errors across the experimental conditions. In short, subjects did not have difficulty perceiving the sentences or encoding them into memory. They did have difficulty, however, in determining whether the sentences

were true or false. This decision required subjects to understand the meaning of the sentences and to respond appropriately to the linguistic content of the message.

The results of the present study suggest that several additional factors related to processing operations involved in language comprehension may be responsible for the observed differences in verification latencies. Because the test sentences were matched on segmental intelligibility, the present findings suggest that, in addition to differences in segmental intelligibility, differences also exist in comprehension between natural speech and high-quality synthetic speech generated by DECtalk and that these differences are above and beyond differences related to the segmental intelligibility of speech, as measured by traditional types of transcription tests or MRT scores (see also Schmidt-Nielsen, 1987; Schmidt-Nielsen & Kallman, 1987). Whatever the precise locus of the differences, the present findings demonstrate that segmental intelligibility alone is not sufficient to account for the pattern of response latencies observed in the present study.

In this connection, it is useful to consider briefly the findings of another recent study carried out by Pisoni & Dedina (1986), who used a sentence verification task with natural speech that had been processed using three quite different digital encoding algorithms. Despite the fact that standard tests of segmental intelligibility using the MRT revealed only very small differences in performance among the three vocoders, the results of the verification task revealed quite robust and consistent findings, which could be related directly to the data rate of the processing algorithms (see also Schmidt-Nielsen & Kallman, 1987). Latencies were fastest and error rates were lowest for the 16 kbps CVSD algorithm, followed by the 9·6 kbps TDHS/SBC algorithm and, finally, the 2·4 kbps LPC-10 algorithm. Thus, it is reasonable to conclude from the present findings using synthetic speech, and the recent data of Pisoni & Dedina using vocoded natural speech, that traditional intelligibility tests may simply be insensitive to important differences that are, in fact, present in the speech waveform. These differences apparently affect the listener's performance in understanding the content of the message and responding appropriately to the truth value of the utterance.

To determine the locus of the observed differences in this study, it is necessary to examine the comprehension process in somewhat greater detail within the framework of a generic model of comprehension. In recent studies of language comprehension, specifically in experiments on sentence verification, it has become common to view comprehension as a "process" and to divide that process into a series of processing stages. Clark & Chase (1972) describe a generic verification model of sentence comprehension with the following four stages: Stage 1 represents the interpretation of the sentence; Stage 2 represents the relevant external or internal evidence; Stage 3 compares the representations from Stages 1 and 2, and Stage 4 responds with the answer computed at Stage 3. According to this model, each stage has one or more cognitive operations and each operation takes some amount of processing time to complete. In applying this model to the sentence verification task used in the present investigation, it is assumed that listeners begin at Stage 1 and by the time they get to Stage 4 they are able to respond with either "true" or "false".

Considering the framework of the verification model outlined above, it is possible to speculate about the locus of the differences observed in the present study. Although we have tried to argue that the differences found in the present study are not due to factors related to segmental intelligibility, and we have been cautious not to over-interpret the results of the present study, it is still possible that our findings are due to some aspect of the perceptual encoding process, either at the time of input or at the time the initial representation of the meaning of the test sentence is constructed at Stage 1 of the model. Thus, the initial representation of the synthetic speech, at Stage 1 in the model outlined above, may be degraded or noisy relative to natural speech. Because standardized tests of speech

intelligibility are not performance limited (i.e., subjects are not typically required to respond rapidly in these tasks), it is quite possible that transcription scores and MRT results typically obtained with high-quality synthetic speech or digitally encoded natural speech are much too insensitive to pick up any differences that are localized at Stage 1 of the verification model at which the initial representation of the sentence is constructed from the speech waveform.

If this line of reasoning is correct, or nearly so, it would imply that the initial representation of the test sentence is encoded in a format that contains some information about the acoustic–phonetic quality or attributes of the input signal. Put another way, some property, or set of properties, related to the perceptual analysis of the speech waveform, and/or its segmental representation, is passed along or "propagated" up the processing system to higher and progressively more abstract levels of the language processing system (see also Streeter & Nigro, 1979). One consequence of this account of our findings would be a general slowing up of all processing activities in the comprehension task under these conditions. This result would not be affected by other experimental manipulations such as sentence predictability or sentence length that may have their effects localized at Stages 2 or 3 in the verification model. Indeed, our failure to find an interaction between voice and predictability would be consistent with this explanation. It would also imply that the locus of the predictability manipulation occurs somewhere later in the comprehension process than the voice manipulation, perhaps at Stage 2, where relevant information is retrieved from long-term memory, or possibly at Stage 3, where the two representations are compared before a response is initiated.

Similar findings using response latencies have been reported by Pisoni (1981) and by Slowiaczek & Pisoni (1982) who used lexical decision and naming tasks to study the perception of isolated words that were either natural speech or synthetic speech generated by the MITalk text-to-speech system. Both studies found longer response latencies for synthetic speech compared to natural speech. However, no interactions were observed with any of the other experimental variables, suggesting, as we found in the present study, that the locus of the effects of the voice manipulation appear to be localized at either the initial stage of perceptual encoding or the development of some initial representation of the input signal that will be used in the comparison process in verification.

Without further studies utilizing additional experimental manipulations, it is not possible to decide which of these alternatives is the correct account of the present results. However, it is clear from the present results that we have found robust effects of the voice manipulation on some selected aspects of the comprehension process that appear to be separable from effects related to overall segmental intelligibility and sentence predictability. The subjects in the present experiment apparently had no difficulty whatsoever in perceiving the words or sentences or responding to the truth value of the meaning of the sentences. Measures of transcription accuracy and verification accuracy were not reliably different for the natural or synthetic speech used here. Our primary finding was that the response latencies were considerably shorter when the sentences were natural speech than when the sentences were produced with high-quality synthetic speech generated by DECtalk.

In summary, the results of the present investigation demonstrate that some aspect of the comprehension process, either the encoding of the initial representation or the comparison process, is affected by the quality of the acoustic–phonetic input in the speech signal. Using short meaningful sentences that were controlled for segmental intelligibility, we found that verification latencies were reliably shorter for natural speech than high-quality synthetic speech produced by rule using DECtalk. Further studies are currently under way in our laboratory to identify the locus of these effects in the human information processing system and to specify the nature of the processing operations that are affected by these differences

in the initial sensory-based input in the speech signal. The results of the present study, taken together with the earlier findings of Manous *et al.* (1985) and the more recent data of Pisoni & Dedina (1986), demonstrate that the sentence verification task appears to be a useful and extremely sensitive method for investigating the comprehension process when initial differences in intelligibility are very small or non-existent. The results also demonstrate the need to develop new and more sophisticated measures of sentence comprehension that can be used to study speech communication at processing levels above and beyond those typically indexed through transcription tasks or traditional forced-choice intelligibility tests, such as the Modified Rhyme Test (MRT) or the Diagnostic Rhyme Test (DRT).

Acknowledgments

This research was supported, in part, by NIH Research Grant NS-12179 to Indiana University in Bloomington. Preparation of this manuscript was also supported, in part, by NSF Grant IRI86-17847. We thank Howard Nusbaum for his helpful comments, suggestions and continuing interest in this research.

References

Allen J. Linguistic-based algorithms offer practical text-to-speech systems. Speech Technology. 1981; 1:12–16.

Allen, J.; Hunnicutt, MS.; Klatt, D. From Text to Speech. Cambridge University Press; Cambridge: 1987.

Chang TM. Semantic memory: Facts and models. Psychological Bulletin. 1986; 99:199–200.

Clark HH, Chase WG. On the process of comparing sentences against pictures. Cognitive Psychology. 1972; 3:472–517.

Clark, HH.; Clark, EV. Psychology and Language. Harcourt Brace Jovanovich; New York: 1977.

Collins AM, Quillian MR. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior. 1969; 8:240–247.

Collins, AM.; Quillian, MR. Facilitating retrieval from semantic memory: The effect of repeating part of an inference. In: Sanders, AF., editor. Attention & Performance III. North Holland Publishing Co; Amsterdam: 1970. p. 304-314.

Conrad C. Cognitive economy in semantic memory. Journal of Experimental Psychology. 1972; 92:149–154.

Egan JP. Articulation testing methods. Laryngoscope. 1948; 58:955–991. [PubMed: 18887435]

Gough PB. Grammatical transformations and speed of understanding. Journal of Verbal Learning and Verbal Behavior. 1965; 4:107–111.

Gough PB. The verification of sentences: The effects of delay of evidence and sentence length. Journal of Verbal Learning and Verbal Behavior. 1966; 5:492–496.

Greene BG, Logan JS, Pisoni DB. Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. Behavior Research Methods, Instruments, & Computers. 1986; 17:100–107.

House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-testing methods: Consonantal differentiation with a closed response set. Journal of the Acoustical Society of America. 1965; 37:158–166. [PubMed: 14265103]

Jenkins, JJ.; Franklin, LD. Recall of passages of synthetic speech. Paper presented at the Psychonomics Society Meeting; November, 1981; 1981.

Larkey LS, Danly M. Fundamental frequency and sentence comprehension. MIT Speech Group Working Papers. 1983; II

Luce PA, Feustel TC, Pisoni DB. Capacity demands in short-term memory for synthetic and natural word lists. Human Factors. 1983; 25:17–32. [PubMed: 6840769]

Manous, LM.; Pisoni, DB.; Dedina, MJ.; Nusbaum, HC. Research on Speech Perception Progress Report No. 11. Speech Research Laboratory, Indiana University; Bloomington, IN: 1985. Comprehension of natural and synthetic speech using a sentence verification task; p. 33-57.

McHugh, A. Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Naval Research Laboratory; 1976. p. 9-9-76.

Moody, TS.; Joost, MG. Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. Proceedings of the Voice Input/Output Society; Alexandria, VA. 1986.

Nusbaum HC, Pisoni DB. Constraints on the perception of synthetic speech generated by rule. Behavior Research Methods, Instruments, & Computers. 1985; 17:235–242.

Nye PW, Gaitenby J. Consonant intelligibility in synthetic speech and in a natural control (Modified Rhyme Test results). Haskins Laboratories Status Report on Speech Research. 1973; SR-33:77–91.

Pisoni DB. Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the Acoustical Society of America. 1987; 70:S98.

Pisoni DB. Perception of speech: The human listener as a cognitive interface. Speech Technology. 1982; 1:10–23.

Pisoni, DB. Some measures of intelligibility and comprehension. In: Allen, J.; Hunnicutt, MS.; Klatt, D., editors. From Text to Speech. Cambridge University Press; Cambridge: 1987. p. 151-171.

Pisoni, DB.; Dedina, MJ. Research on Speech Perception Progress Report No. 12. Speech Research Laboratory, Indiana University; Bloomington, IN: 1986. Comprehension of digitally encoded natural speech using a sentence verification task (SVT): A first report; p. 3-18.

Pisoni, DB.; Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted test-to-speech system. 1980 IEEE International Conference Record on Acoustics, Speech, and Signal Processing; IEEE Press, New York. 1980. p. 572-575.

Pisoni DB, Nusbaum HC, Greene BG. Perception of synthetic speech generated by rule. Proceedings of the IEEE. 1985; 11:1665–1676.

Schmidt-Nielsen, A. Evaluating degraded speech: Intelligibility tests are not all alike. Proceedings of Military Speech Tech; New York: Media Dimensions; 1987. p. 118-121.

Schmidt-Nielsen, A.; Kallman, HJ. NRL Report No. 9089. Washington, D.C: Naval Research Laboratory; 1987. Evaluating the performance of the LPC 2.4 kbps processor with bit errors using a sentence verification task.

Schwab EC, Nusbaum HC, Pisoni DB. Some effects of training on the perception of synthetic speech. Human Factors. 1985; 27:395–408. [PubMed: 2936671]

Slowiaczek LM, Pisoni DB. Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America. 1982; 71:S95–96.

Streeter LA, Nigro GN. The role of medial consonant transitions in word perception. Journal of the Acoustical Society of America. 1979; 65(6):1533–1541. [PubMed: 489823]

Appendix. Stimulus materials used in sentence verification test

A. Three-word, false, high-predictability sentences

1. Men wear dresses.
2. Circles are square.
3. Sandpaper is smooth.
4. Winter is hot.
5. Screaming is soft.
6. Ice is hot.
7. Fire is cold.
8. Skyscrapers are short.
9. Marshmallows are hard.
10. Pillows are hard.

B. Three-word, false, low-predictability sentences

1. Coffee is hard.
2. Roses are extinct.
3. Money buys peace.
4. Diamonds are soft.
5. Jails are playgrounds.
6. Coffee is salty.
7. Clocks run laps.
8. Candy is sugarless.
9. Snakes can sing.
10. Crackers are juicy.

C. Three-word, true, high-predictability sentences

1. Books have pages.
2. Honey is sweet.
3. Fire can burn.
4. Scissors cut paper.
5. Birds have wings.
6. Babies often cry.
7. Coal is black.
8. Babies wear diapers.
9. Headaches are painful.
10. Doctors prescribe medicine.

D. Three-word, true, low-predictability sentences

1. Dogs are mammals.
2. Heat melts polyester.
3. Cats have legs.
4. Oranges are nutritious.
5. Games are played.
6. Giraffes are spotted.
7. Frisbees are objects.
8. Bakers make pastries.
9. Bacon is food.
10. Leopards have tails.

E. Six-word, false, high-predictability sentences

1. People in libraries should talk loudly.

2. Doctors try to make people sick.
3. Most flowers bloom in the winter.
4. Sunglasses are most useful at night.
5. A ZIP code has five letters.
6. Poor people have lots of money.
7. Sleeping pills will keep you awake.
8. Babies cry when they are happy.
9. China plates are hard to break.
10. The earth revolves around the moon.

F. Six-word, false, low-predictability sentences

1. One plus one will make ten.
2. Fathers are younger than their daughters.
3. There are seven days per semester.
4. Wives often divorce their own brothers.
5. Children like candy and getting drowned.
6. Most people sleep in their kitchen.
7. Newborn babies know how to spell.
8. Florida is located in the Arctic.
9. The color of clouds is magenta.
10. Leap year comes every four minutes.

G. Six-word, true, high-predictability sentences

1. Pots and pans are used for cooking.
2. When peeling onions people often cry.
3. We cut steak with a knife.
4. Large cuts may often require stitches.
5. A sandwich is made with bread.
6. Hospitals are filled with sick people.
7. Flowers have leaves and pretty petals.
8. One hundred pennies equal a dollar.
9. A blanket will keep you warm.
10. The stars shine in the sky.

H. Six-word, true, low-predictability sentences

1. Fish can swim but can’t smoke.
2. Smoking is bad for your teeth.
3. Our alphabet has twenty-six characters.

4. A triangle has only three vertices.
5. Hawaii's a good place to sunbathe.
6. You boil water to make rice.
7. Most people like to receive love.
8. Seat belts are worn for driving.
9. The sun is a fiery sphere.
10. Riding a horse can be bouncy.

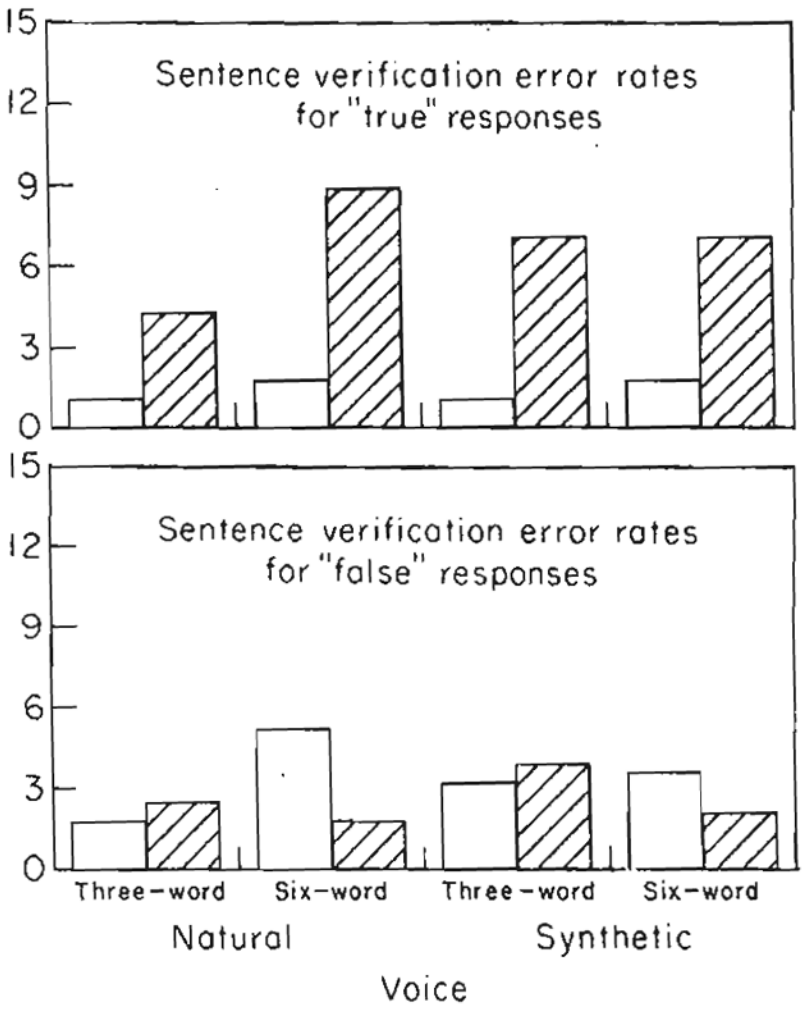

Figure 1.
Sentence verification error rates for “True” responses (top panel) and “False” responses (bottom panel) for natural and synthetic speech at each of two sentence lengths. ☐ High Predictability. ▨ Low Predictability.

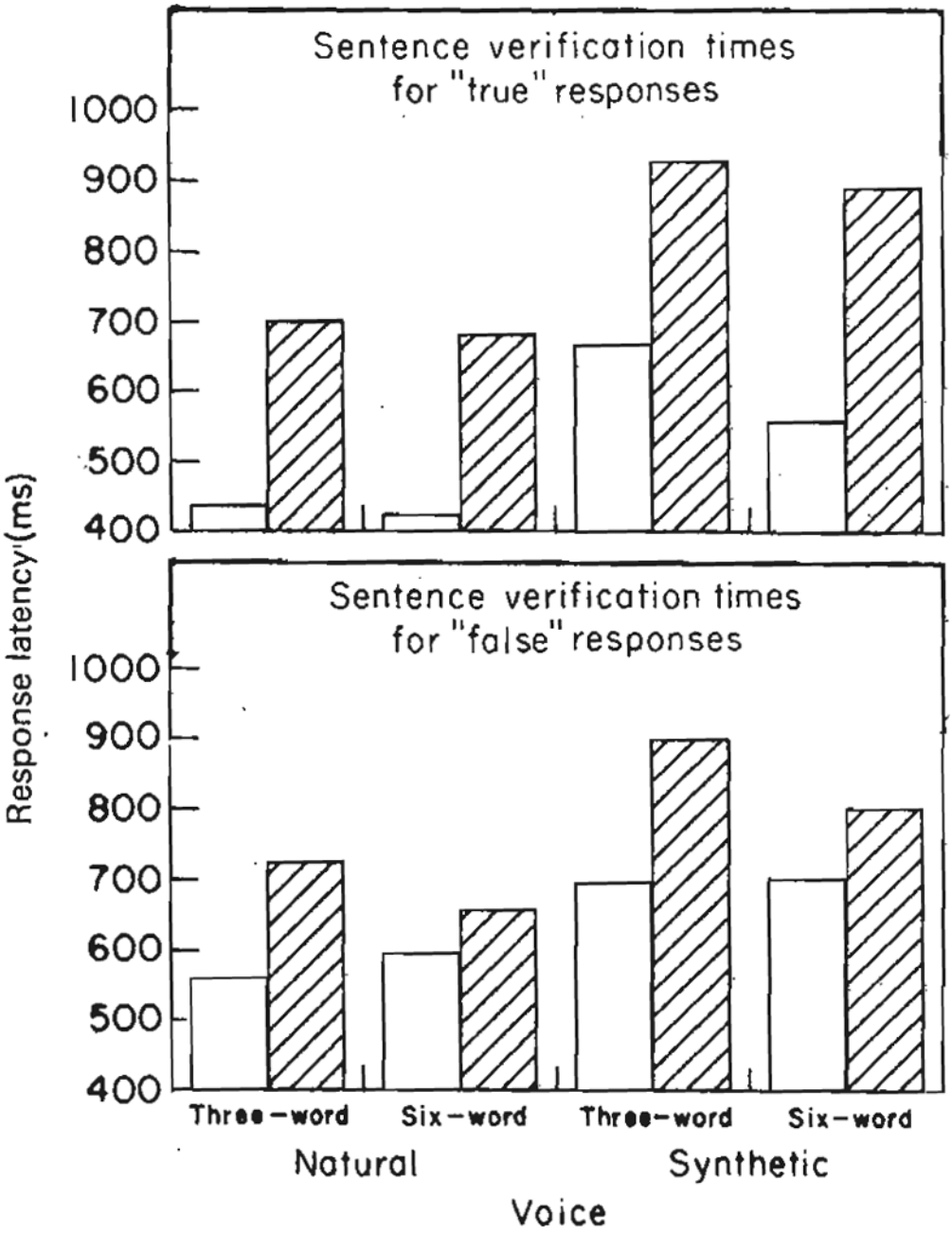

Figure 2.

Mean sentence verification latencies (in ms) for “True” responses (top panel) and “False” responses (bottom panel) for natural and synthetic speech at each of two sentence lengths. The high-predictability sentences are displayed with open bars; the low-predictability sentences are displayed with striped bars. The response latencies shown in this figure are based on only those trials in which subjects both verified *and* transcribed the sentence correctly.

☐ High Predictability. ▨ Low Predictability.

Table I

Examples of test sentences used for predictability norms

Three-word sentences
Cotton is _____. Birds can _____.
Six-word sentences:
Pots and pans are used for _____. Most businessmen wear suits to _____.

Table II

Examples of high- and low-predictability sentences

High-predictability sentences
Three-word → Giraffes are *tall.* (True)
Sandpaper is *smooth.* {False)
Six-word → Pots and pans are used for *cooking.* (True)
Sunglasses are most useful at *night.* (False)
Low-predictability sentences
Three-word → Trees have *greenery.* (True)
Diamonds are *rough.* (False)
Six-word → Most businessmen wear suits to *lunch.* (True)
Leap year comes every four *minutes* (False)

Table III

Mean response latencies in ms for predictability and length

	True	
	High predictability	**Low predictability**
Three-word	553	815
Six-word	492	788
	False	
	High predictability	Low predictability
Three-word	629	813
Six-word	650	730

Table IV

Mean response latencies in ms for voice and length

	True	
	Natural	**Synthetic**
Three-word	570	799
Six-word	555	725
	False	
	Natural	Synthetic
Three-word	643	799
Six-word	627	753