Published in final edited form as:

Lang Speech. 1992 ; 35(Pt 4): 351–389.

COMPREHENSION OF SYNTHETIC SPEECH PRODUCED BY RULE: A REVIEW AND THEORETICAL INTERPRETATION*

Susan A. Duffy and
Amherst College

David B. Pisoni
Indiana University

Abstract

In this paper, we review research on the perception and comprehension of synthetic speech produced by rule. We discuss the difficulties that synthetic speech causes for the listener and the evidence that the immediate result of those difficulties is a delay in the point at which words are recognized. We then argue that this delay in processing affects not only lexical access but also comprehension processes. We consider the mechanisms by which the comprehension system adjusts to this delay, the resulting costs to higher level comprehension processes, and the changes that occur in the language processing system as its familiarity with synthetic speech increases. Based on the framework we have developed, we suggest several directions for future research on the comprehension of synthetic speech.

Keywords

synthetic speech; intelligibility; comprehension; word recognition

Introduction

The use of synthetic speech in a variety of communication settings has been growing rapidly over the last ten years. Although early research on the perception of synthetic speech focused on evaluating the intelligibility of individual phonemes and words in isolation, more recent research efforts have focused on understanding how human listeners process synthetic speech to the point of comprehension, and how that processing might differ from that of natural speech. In this paper, we review and interpret the existing literature in order to learn more about how synthetic speech affects the language comprehension system (LCS) as a whole. We feel that such an endeavor will help us to better understand the LCS itself. This paper has three general goals: First, we want to bring together and summarize the literature on the perception and comprehension of synthetic speech. Second, we want to use this literature to understand the effects of degraded acoustic input on the LCS. And third, we want to develop a theoretical framework that will guide future research efforts in this area.

*This paper was written while the first author was on sabbatical leave at the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. Preparation of this paper was supported, in part, by NSF Grant IRI-86-17847, and, in part, by NIDCD Research Grant DC-00111-15 to Indiana University. Additional support was provided by NIMH Grant 5 RO3 MH46855 to Amherst College. We thank Ellen Gurman Bard, Steven Goldinger, Sheri Hunnicutt, Lynne Nygaard, and Bruno Repp for thoughtful reviews of earlier versions of the paper.

Copyright (c) 2007 ProQuest LLC, Copyright (c) Kingston Press

Correspondence regarding this paper should be addressed to Susan Duffy, Department of Psychology, Amherst College, Amherst, MA 01002 (SADUFFY@AMHERST.edu).

We can think of the LCS as a highly complex and efficient system that is optimally designed to extract meaning in real time from acoustic signals generated by a human vocal tract. That is, the input is speech sounds and the output is a representation of the meaning of the utterance. What happens between input and output is, of course, a current topic of much debate (see, for example, papers in Altmann, 1990; Frauenfelder and Tyler, 1987; Marslen-Wilson, 1989). The question we want to focus on here is how processing differs when the input to the LCS is synthetic speech produced automatically by rule. Presumably, processing at the input end of the system becomes more difficult when the input is synthetic speech. It is less clear how later comprehension processes might be affected by the change in the input signal.

Numerous studies indicate that human listeners are capable of comprehending synthetic speech and of doing so in real time (e.g., Ralston, Pisoni, Lively, Greene, and Mullennix, 1991). Indeed, one of the striking results in many of the early studies carried out on the comprehension of synthetic speech was the finding that large differences in the segmental intelligibility of synthetic speech compared to natural speech were accompanied by only small or negligible differences in overall scores on a set of multiple-choice comprehension questions (e.g., Ingemann, 1978; Jenkins and Franklin, 1982; McHugh, 1976; Nye, Ingemann, and Donald, 1975; Pisoni and Hunnicutt, 1980). We will argue, however, that the finding of comparable levels of comprehension for synthetic and natural speech is not necessarily evidence that the LCS operates in the same way when processing synthetic and natural speech. In fact, we will argue just the opposite: Recent findings provide support for the claim that human listeners must adjust in subtle but systematic ways to cope with synthetic speech as an input signal. Furthermore, there are processing costs associated with the comprehension of synthetic speech, costs that become apparent only when appropriate dependent measures are used to index processing load.

The present paper is organized into seven sections. In the first section, we outline a set of assumptions that we have made about the LCS as a whole. In the second section, we briefly discuss how synthetic speech differs from natural speech. In the third, we identify the immediate effect of such differences on low-level perceptual processes that are focused on simply identifying the phonemes and words in the sound stream. In the next three sections, we go on to discuss how these perceptual differences affect the comprehension system as a whole. Here we focus on immediate adjustments, resulting costs, and long-term changes, respectively. Finally, in the last section, we draw some general conclusions from our review of the literature and speculate about future directions for research on the comprehension of synthetic speech produced by rule.

A Model of Spoken Language Comprehension

We begin by considering three general assumptions about the LCS for which there is some consensus in the field. First, we assume that the LCS is composed of several subsystems, or processors, each focused on a different task within the comprehension process. Although the specific processors may differ from model to model, in general, processors can be divided into two kinds: One set of processors has as a final goal the identification of words in the discourse (e.g., a phonetic processor, a lexical access processor). A second set of processors is devoted to combining words into larger representations of structure and/or meaning (e.g., syntactic and message level processors). We will refer to the former processes as “lower-level” and to the latter as “higher-level comprehension” processes. See Just and Carpenter (1987), and papers in Balota, Flores d’Arcais, and Rayner (1990), Altmann (1990), and Coltheart (1987) for more detailed discussion of various language processors relevant to spoken and written language comprehension.

A second assumption is that although these processors may operate in parallel, there are constraints on the ordering of various processes. For example, the message level processor cannot integrate a given word into the overall message until the lexical processor has actually identified that word. This does not, however, imply that all message level processing occurs after all lexical access processing is completed. The message level processor can integrate all words that have been identified at a certain point while the lexical processor is working on the next word (Duffy and Rayner, 1990). Thus, this kind of model has both serial and parallel characteristics (see Thibadeau, Just, and Carpenter, 1982).

A third assumption is that limited computational resources are available for carrying out all of these comprehension processes. This resource assumption has been central to many current cognitive models (e.g., Schneider and Shiffrin, 1977; Kahneman, 1973). Each of the processes involved in comprehension can draw upon limited attentional resources. Processes that are well-practiced may require very few resources; these processes are sometimes referred to as automatic. Processes that are most likely to require minimal resources include processes involved in recognizing phonemes and words. Processes that are assumed to require many resources are the “higher-level” processes involved in comprehension. A number of models of comprehension now detail such higher-level processes. For example, information from earlier in the text must be reinstated into the working memory buffer for comprehension of the current sentence (Kintsch and van Dijk, 1978; McKoon and Ratcliff, 1980; O’Brien, Duffy, and Myers, 1986). Causal inferences based on world knowledge must be generated to link sentences in the text (Myers, Shinjo, and Duffy, 1987). Complex representational structures must be built, involving time consuming shifts from one substructure to another as topics shift within the text, and activated information that is not currently relevant must be suppressed (Gernsbacher and Faust, 1991; Gernsbacher, Varner, and Faust, 1990).

Two current models from the reading literature are especially relevant to our interpretation of the literature on the perception and comprehension of synthetic speech. These are Stanovich’s interactive-compensatory model (Stanovich, 1980) and Perfetti’s verbal efficiency model (Perfetti, 1985). At the heart of both of these models is the claim that when “lower-level” word recognition processes are not highly automatized (as is the case for poor readers), their operation drains needed resources away from the higher-level comprehension processes that use them. We will argue that a parallel analysis can be made for the perception of synthetic speech: Lower level phoneme- and word-recognition processes require more resources than usual for identifying words in synthetic speech. The result is a drain of resources away from higher level comprehension processes. Thus, the effect of difficulty at the word recognition level is not confined only to that level but affects higher level comprehension processes as well.

Differences in Perception Between Natural and Synthetic Speech

We turn now to a review of research on the perception of synthetic speech and the problems it poses for the phonetic processor. Although systems that generate synthetic speech have improved substantially over the last decade, synthetic speech still differs in several fundamental ways from natural speech (see Klatt, 1987). At a gross level, listeners have no trouble determining that they are listening to synthetic speech because it still has a distinctive mechanical sound to it. Listeners frequently report that the speech sounds like a computer. The differences in perception between synthetic and natural speech seem, at first glance, to be associated with the lack of appropriate prosody in the utterances and with stereotyped coarticulation routines used for segmental synthesis.

At a more fine-grained level, however, synthetic speech differs from natural speech in terms of its dynamic acoustic-phonetic structure. In natural speech, individual phonemes are specified redundantly with multiple acoustic cues to each segmental contrast. This redundancy typically does not exist in synthetic speech because only a small set of acoustic cues have been implemented for each phoneme in the rule systems used for segmental synthesis (Klatt, 1987). The goal has been to use the minimal set of cues necessary for specifying a given phoneme (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952; Cooper, Liberman, and Borst, 1951; Liberman, Ingemann, Lisker, Delattre, and Cooper, 1959). As a result, the acoustic-phonetic information lacks redundancy; in some cases, a given phonetic distinction may be signalled by only one acoustic-phonetic cue. Thus, if the listener misperceives a particular cue, other redundant cues are not available to support perception of the intended phoneme in a particular context.

Clark, Dermody, and Palethorpe (1985) provide evidence to support this characterization of synthetic speech. In their experiment, subjects heard natural and synthetic CV syllables and were asked to identify the syllable. Syllables were presented in noise; the noise was adjusted to yield an identification accuracy score of about 50% on the first trial. Syllables were presented five times, and accuracy across the set of presentations was recorded. Their results indicated that repetitions helped in the identification of natural speech but did not help in the identification of synthetic speech. Presumably, repetitions facilitated identification of natural speech syllables because the listener could pick up additional, redundant cues on successive trials until a phoneme became intelligible. The authors argue that repetitions failed to improve the identification of synthetic syllables because the additional redundant cues were not present in the acoustic signal.

Nusbaum, Dedina, and Pisoni (1984) investigated another possible difference between natural and synthetic speech resulting from the lack of redundancy. The minimal cues available for a given phoneme may, in some contexts, actually provide misleading or inappropriate information. This view leads to two kinds of predictions. For some phonemes, perception of synthetic speech will be similar to natural speech presented in noise, in which some of the redundant cues are masked. Thus, we would expect to find similar distributions of confusion errors for synthetic speech and natural speech in noise. For other phonemes, however, little overlap would be expected in the distributions of confusion errors for the two types of speech. These are the phonemes for which the acoustic cues are inappropriate and misleading.

In order to evaluate these predictions, Nusbaum *et al.* (1984) examined the distributions of confusions for phonemes in consonant-vowel syllables using three synthesizers as well as natural speech presented at several signal-to-noise levels. For some phonemes, the distributions of errors for synthetic speech overlapped closely with those for natural speech. For example, Votrax stop consonants tended to be confused with other stop consonants, a confusion pattern similar to that for natural speech. But for other phonemes, there were striking differences between the confusion patterns for synthetic and natural speech. For example, Prose-2020 stop consonants tended to be confused with fricatives. When the DECTalk /r/ was misidentified, it was always misidentified as /b/. These confusion patterns did not resemble those for natural speech. Thus, Nusbaum *et al.* (1984) characterized synthetic speech as having an acoustic cue structure that is, at times, misleading as well as, at the same time, lacking in the acoustic-phonetic redundancy typically found in natural speech

Perception of Synthetic Words in Isolation

The missing or misleading acoustic-phonetic cues in synthetic speech, of course, have perceptual consequences for the listener. Many behavioral studies have shown that it is much more difficult to identify phonemes and hence words in synthetic speech compared with natural speech. This difficulty is reflected directly in measures of identification accuracy and response time, and indirectly in measures of concurrent processing load.

Identification accuracy

In early work on the perception of synthetic speech, the main evaluative tools were behavioral tests that measured identification accuracy. In these experiments, subjects were presented with a particular stimulus and were required to identify it from a closed response set. Traditionally, within this literature, identification accuracy has been called “intelligibility”.

In the first study of this kind using synthetic speech, Nye and Gaitenby (1973) employed the Modified Rhyme Test (MRT) to assess the intelligibility of consonants produced by rule using the Haskins speech synthesizer. In the MRT task, subjects are presented with a CVC word and then must select the correct item from a set of six alternatives. For example, subjects hearing “tame” have the six responses: *game*, *fame*, *tame*, *came*, *name*, and *same* to choose from. Nye and Gaitenby found error rates of 2.7% for natural speech and 7.6% for synthetic speech.

The MRT has become a standard for testing the intelligibility of synthetic speech systems (Nye and Gaitenby, 1973; Pisoni and Hunnicutt, 1980; Logan, Greene and Pisoni, 1989; Allen, Hunnicutt and Klatt, 1987). In a test of seven text-to-speech synthesis systems, Logan *et al.* (1989) found error rates for synthetic speech ranging from 3% to 36% compared to 0.5% for natural speech. These results are shown in Figure 1.

Also included in Figure 1 are results from an open response version of the MRT. In this procedure subjects are not forced to choose a response from among a set of six alternatives. Instead, they simply write down the word they hear on each trial. Performance on the open format MRT is uniformly lower for natural speech as well as for all of the speech synthesizers tested, although the drop in performance is much larger for the synthetic speech than the natural speech. The overall error rate on the open response test for natural speech was 2.8%; the error rate for the best synthesizer (DECTalk Paul) was 12.9%; error rates for the worst synthesizers were above 50%.

Two important conclusions can be drawn from the Logan *et al.* intelligibility data. First, it is clear that even the best speech synthesizers are not comparable in segmental intelligibility to natural speech under the same testing conditions. Second, the range of intelligibility levels for different kinds of synthetic speech is substantial. As we review the literature on the perception of synthetic speech, it is important to keep in mind that results for the same tasks may differ for different kinds of synthetic speech just as the pattern of results for synthetic speech may differ from those obtained with natural speech.

Identification time

Even in those cases in which natural speech and synthetic speech can be equated for identification accuracy, several studies have shown that differences arise in the time it takes to process a string of phonemes. In an early study, Pisoni (1981) used a lexical decision task to measure recognition times for natural and synthetic words and non-words. Mean response times and percent correct scores are presented in Figure 2. As expected, accuracy was greater for natural speech, mirroring the intelligibility results summarized above. In

addition, response times for those items responded to correctly were longer for the synthetic stimuli by 145 msec. This increase in response time was observed for non-words as well as words. The parallel differences in response times for natural and synthetic speech for non-words suggests that the extra processing time is spent in phoneme identification rather than in locating items in the lexicon.

Pisoni (1981) also carried out a naming task using the same stimuli. Subjects heard a target item (either a word or a non-word) and simply had to repeat it aloud as quickly as possible. Again, response times were longer for the synthetic stimuli; this effect held for both words and non-words, suggesting a difference in phonetic coding at early levels of perceptual analysis.

Finally, processing time differences have also been found in experiments using the gating paradigm. In natural speech, listeners do not always need to hear the whole word in order to identify it (Grosjean, 1980). For example, Salasoo and Pisoni (1985) found that subjects needed on average only the first 56% of a content word in a semantically coherent sentence to identify it accurately in context. This percentage increased to 72% for content words in anomalous sentences and 82% for the same words excised from sentences and presented in isolation.

Using a gating paradigm, Manous and Pisoni (1984) measured the amount of stimulus duration needed to identify isolated natural and synthetic words that were equated for length. They used a high-quality synthesizer, DECTalk. In their task, subjects heard successive presentations of a given target word presented in isolation. On the first presentation, the subject heard the first 50 msec of the word; on the second presentation the subject heard the first 100 msec; on the third, the first 150 msec, and so on until the whole word had been presented. After each presentation, subjects were asked to write down the word they thought they heard. The dependent variable of interest was the duration of the stimulus at which a subject could identify the word correctly.

The results indicated that for natural speech, subjects needed on average 361 msec of a target word to identify it correctly. In contrast, for synthetic speech, subjects needed 417 msec of the same target word. There are at least two reasons why such a difference might occur. First, there may be a minimal number of phonemes needed to identify a given word; because subjects cannot readily identify all phonemes of a synthetic word, they may need to hear more phonemes to reach a threshold for recognition. Second, subjects may need more of the surrounding acoustic context to help determine what a particular phoneme is in synthetic speech than in natural speech. Thus, more of a synthetic word must be heard in order to identify the same initial set of phonemes identified for the natural word.

The studies reviewed in this section suggest that at least two factors may cause a delay in word recognition for synthetic speech even when identification accuracy is high. First, the gating study demonstrated that for synthetic speech listeners need to hear more phonemes of a given word in order to successfully identify it. The longer response times observed for synthetic speech in the lexical decision and naming tasks may reflect this phenomenon. In addition, however, these differences may also reflect the contribution of a second factor. It may simply take longer for listeners to identify a given phoneme in synthetic speech than natural speech because there is less redundancy in the acoustic-phonetic specification of a phoneme. The important point however, is that, in isolation, synthetically produced words require more processing time than naturally produced tokens of the same items.

Concurrent load measures

Given that synthetic words require more processing resources (in terms of time), it should also be the case that the identification of synthetically produced words should leave fewer processing resources for carrying out other tasks. Exactly this point has been made in a series of studies focusing on the recall of spoken word lists in which the demands on working memory were varied.

In the first study, Luce, Feustel, and Pisoni (1983) had subjects recall lists of 15 words that were either natural speech tokens or high-quality synthetic (MITalk) speech. Before hearing the list of spoken words, some subjects were given a list of three or six visually-presented digits to remember as a memory pre-load. After the word list was presented, subjects were required to write down the digit list in order and then recall as many of the spoken words as they could. The word recall results are presented in Figure 3.

As expected, significant effects of voice and load were observed in the word recall data. Listeners recalled fewer words from the synthetic lists. Listeners also recalled fewer words when they had more digits to remember. However, the two factors did not interact. The expected interaction (although a weak one) was found only in the digit recall data which are presented in Figure 4. The increase in memory load caused a larger drop in performance on the digit recall task paired with the synthetic word lists than with the natural word lists.

In a subsequent experiment, Luce *et al.* (1983) tried a different manipulation to find stronger evidence for the increased capacity demands of synthetic speech. In this experiment, subjects were asked to recall lists of 10 words; furthermore, they were instructed to recall the words in serial order. Luce *et al.* argued that recalling words in order requires more memory capacity because information about serial position must be stored along with information about the identity of each item on the list.

The resulting serial position curves are presented in Figure 5. More words were recalled from the natural lists across all serial positions than from the synthetic lists. In addition, recall was better for items from the primacy and recency portions of the serial position curve (Murdock, 1962). However, the primacy effect was not as great for synthetic speech as it was for natural speech. That is, the recall advantage typically found for words presented at the beginning of a list compared with words in the middle of the list was much smaller for synthetic speech than for natural speech. Although the main effect of voice can be accounted for by differences in identification accuracy between natural and synthetic speech, the interaction with serial position cannot. Luce *et al.* argued that the reduced primacy effect for the synthetic lists reflects a reduced capacity for transferring information from working memory into long-term memory (see also Goldinger, Pisoni, and Logan, 1991; Martin, Mullennix, Pisoni, and Summers, 1989; Rundus, 1971).

In a follow-up study, Luce and Pisoni (1983) replicated the serial recall study with synthetic and natural lists for which transcription data on individual words yielded at least 98% accuracy. Again, subjects recalled more words from the natural lists than from the synthetic lists. Furthermore, the synthetic-natural difference was again larger in the primacy section than in the middle or recency portion of the serial position curve. In contrast to the earlier study by Luce *et al.* (1983), the main effect of voice observed here cannot be accounted for by recourse to differences in identification accuracy. The results can be accounted for, however, in terms of the recognition time differences outlined earlier. More resources (in terms of processing time) are needed to recognize a given synthetic word, leaving fewer resources available for encoding the serial position and for transferring item and order information into long-term memory.

Summary

When synthetically produced words are presented in isolation, listeners have trouble identifying the constituent phonemes. This difficulty shows up in accuracy measures, processing time measures, and other measures of processing load. In some cases, the listener cannot identify a phoneme at all; in other cases, identification is possible but the process requires more time. If one considers the LCS to be composed of a series of modules or processors (e.g., Fodor, 1983; Forster, 1979), then the immediate effect of these differences should be on the phonetic processor. The differences in processing time and accuracy reflect difficulty in constructing the initial phonetic representation from the speech waveform. The result is an inability to identify a particular word or a delay in the point at which a word is recognized. In the next section, we consider how the LCS as a whole may be affected when the phonetic processor encounters difficulty in recognizing phonemes.

Adjustments in the Language Comprehension System

The phonetic processor is, of course, part of a more complex system for comprehending spoken language. In the experiments discussed above, this processor was disengaged, so to speak, from the rest of the system, because the test words (and nonwords) were presented in isolation dissociated from the discourse context. In this section, we examine what happens when words are presented in discourse context and the full comprehension system is engaged.

The perceptual analyzers that contribute to spoken word recognition can be thought of as a set of processors that operate under strict deadlines. Given the brief duration of any word in working memory, it is important for word recognition to be achieved before subsequent utterances overwrite the current trace in memory. We have already seen that, in isolation, the point at which word recognition is achieved occurs later in time for synthetic speech than natural speech. In this section, we consider the evidence that the comprehension system as a whole has mechanisms that compensate for problems encountered by the perceptual processor, mechanisms that help to minimize word recognition time and insure that comprehension can take place.

Use of context

The most obvious compensatory mechanism is to use the surrounding semantic and syntactic context to speed word recognition and increase the probability of its success. Such a use of context has been observed in classic studies of natural speech presented in noise (e.g., Miller, Heise, and Lichten, 1951). A number of studies provide strong evidence that context plays a much greater role in the identification of synthetically produced words than naturally produced words.

Mirenda and Beukelman (1987) directly compared transcription accuracy for natural and synthetic words in isolation and in sentences. For natural speech, there was no difference between word and sentence accuracy (both were 99%). In contrast, for the best synthetic speech (DECTalk Betty) accuracy was 82% on word transcription but 96% on sentence transcription. An even larger effect was observed for Votrax (50% accuracy on words but 76% accuracy on sentences). The differences disappeared, however, for the worst of the synthesizers (Echo) for which word accuracy (35%) was only 4% worse than sentence accuracy (39%). This latter finding, which probably reflects a floor effect, suggests that context can be used as long as some part of the context is itself intelligible. In another study, Hoover, Reichle, Van Tasell, and Cole (1987) found a similar kind of interaction in which performance was equivalent for Echo and Votrax when single words were presented, but Votrax outperformed Echo when sentences were presented. Hoover *et al.* also found that

even a neutral context (“Say the word ...”) facilitated transcription performance for synthetic speech.

Comparable effects of context have been observed in the comparison of transcription accuracy for semantically meaningful sentences and semantically anomalous sentences. The semantically anomalous sentences were originally developed at Haskins Laboratories (Nye and Gaitenby, 1974) and were used in evaluating their synthesis system. Pisoni and Hunnicutt (1980) examined transcription accuracy for meaningful and anomalous sentences for natural and synthetic speech. They found, again, little opportunity for context to contribute to performance for natural speech; accuracy on the anomalous and meaningful sentences was at ceiling (anomalous: 97.7% correct, meaningful: 99.2% correct). However, context did facilitate performance for synthetic speech (anomalous: 78.7%, meaningful: 93.2%). Again, for high-quality synthetic speech, the addition of context dramatically reduced the differences in performance between natural and synthetic speech. It is as if context can be used to compensate for poor or ambiguous acoustic-phonetic structure, a finding that is not surprising given what is known about the role of context in natural speech intelligibility studies (e.g., Miller *et al.*, 1951).

Finally, Simpson (1975) examined the effects of a different kind of context — prior knowledge of a specialized vocabulary — on transcription of natural and Votrax synthetic speech. Her subjects were pilots and non-pilots who listened to aircraft messages. She found that pilots (the experts) were more accurate than non-pilots; subjects were also more accurate in the natural voice condition. Furthermore, expertise and voice interacted. The beneficial effect of expertise was much larger for synthetic speech than for natural speech. Again, knowledge that allows the listener to narrow down the possibilities is much more useful in understanding synthetic speech because of the degraded or misleading acoustic-phonetic information in the signal (Miller *et al.*, 1951; Pollack, 1959, 1964).

Within these transcription tasks, the evidence suggests a greater contribution of top-down processing in the perception of synthetic speech compared with natural speech. Presented with synthetic speech, the language processing system uses semantic context and prior knowledge of specialized vocabulary (what Simpson and Navarro, 1984, label “intelligibility enabling factors”) to aid it in identifying words. This seems to be a reasonable way for a system to compensate for difficulties in reliably identifying phonemes that are encoded poorly in the acoustic signal.

This account is similar to views proposed by the interactive-compensatory models that have been developed in the field of reading (e.g., Stanovich, 1980; Perfetti, 1985). These models were formulated to account for individual differences in visual processing between good and poor readers. Within these models it is assumed that poor readers compensate for suboptimal word recognition skills through the use of semantic context (West and Stanovich, 1978; Perfetti and Roth, 1981). Thus, when the comprehension system is faced with difficulties in low-level perceptual analyses (analysis of synthetic phonemes for listeners faced with synthetic speech or analysis of visual information for poor readers) one immediate strategy is to rely much more on semantic context. A system that is strongly bottom up and highly automatized in its skilled form (e.g., the system of the skilled reader or listener) becomes one that is more top down and strategically oriented. One might speculate that similar kinds of changes would occur for listeners hearing speech produced with a foreign accent or speech spoken by a deaf person or an individual with a speech impediment or a young child.

It should be noted here that there is currently some disagreement within the speech perception literature on the extent to which context can and does directly become involved in phoneme recognition (e.g., Ganong, 1980; Marslen-Wilson and Welsh, 1978; Samuel,

1981). We do not mean to claim that context is directly involved in the perceptual analysis of phonemes from the speech waveform. Rather, the claim here is a more general one that context is used somewhere in the processing system to facilitate the identification of synthetically produced words. Furthermore, context appears to be used by the listener to compensate for impoverished acoustic-phonetic information in the signal.

In the studies we have discussed so far, the evidence regarding the use of context in the comprehension of synthetic speech has been drawn from accuracy data. The results indicate that context helps when a word is otherwise unidentifiable based on its acoustic-phonetic properties. Because words can be identified almost perfectly in the read natural speech that is used in these studies, effects of context cannot be observed in the identification accuracy data because of ceiling effects. One would like to study the relative effects of context on measures for which natural speech may not be at ceiling, such as response time. We know of no studies of the relative effects of context on lexical decision or naming time for natural and synthetic speech. However, we have recently completed a gating study to assess the effects of sentence context on the signal duration needed for recognition of natural and synthetic words (Duffy and Pisoni, 1991).

In this study, subjects heard a sentence context followed by a target noun that was gated over a range of signal durations. Sentence contexts were either semantically congruent or neutral with respect to the target word. Stimuli were taken from a set used by Stanovich and West (1981) and modified by Duffy, Henderson and Morris (1989) to study context effects on visual word recognition. Examples of the stimuli are given below (with the target word underlined):

Congruent: The soldiers flew in the helicopter.

Neutral: The men stood near the helicopter.

Subjects heard either naturally produced or synthetically produced versions of these sentences. The duration of each gate was 50 msec. We analyzed the data in two ways. First, we computed the mean signal duration needed to recognize a word. For this analysis, we excluded all those trials on which a subject never identified the target word; thus, we included only those trials in which the word was ultimately correctly identified. Data are presented in Table 1.

We found that subjects needed less signal duration to identify natural words than synthetic words (replicating the earlier results of Manous and Pisoni, 1984). We also found that subjects needed less signal duration to identify words in the congruent contexts than words in the incongruent contexts. Context and voice did not interact. That is, in contrast to the transcription studies discussed earlier, context had the same effect for both natural and synthetic speech. It is clear from these data that, in the neutral condition, performance for the natural speech was not at ceiling. Thus, there was room for context to have some effect on recognition performance.

In a second analysis, we simply computed the percentage of trials on which a word was never correctly identified; these are the trials that were excluded from the previous analysis. These percentages are presented in parentheses in Table 1. Here, we do find an interaction of voice and context. Subjects made many more errors in the synthetic-neutral condition than in the other three conditions of the experiment.

Taken together, these results suggest that when the acoustic-phonetic information in a synthetic speech signal is so inadequate that it does not alone permit word identification, the LCS uses context in a compensatory way. There is no evidence, however, that the use of context can eliminate the recognition time differences in synthetic and natural speech. That

is, even though context can reduce the signal duration needed to identify a synthetic word, it does so to the same degree for natural speech as well. Although we would like to see converging evidence from response time studies of context effects, the results of this gating study suggest that even in appropriate context, synthetically produced words reach the point of recognition later in the utterance than naturally produced words. This finding has implications for the relative timing of comprehension processes that must await the completion of word recognition before they can be carried out.

Attention to surface details

The research on transcription accuracy overwhelmingly suggests a greater dependency on semantic context for identifying words in synthetic speech compared with natural speech. The results of studies on comprehension of long passages of fluent speech by Luce (1981) and by Ralston *et al.* (1991) suggest another difference in the processing of synthetic and natural speech at the word level. These studies found that listeners paid more attention to surface information in synthetic speech than in natural speech.

In a comprehension study, Luce used a true/false recognition test to probe memory for several kinds of information within his texts: propositions of high importance, propositions of low importance, inferences, and surface (word) information. He found that synthetic speech was worse than natural speech on the measures that directly assessed meaning (propositions and inferences). However, synthetic speech was *better* than natural speech on the measure that assessed memory for exact words used in the text. These results are presented in Figure 6.

In a more recent study, Ralston *et al.* (1991) also tested recognition memory for propositions and individual words from a series of natural and synthetic texts that subjects heard. Their results are presented in Figure 7. In contrast to Luce (1981), Ralston *et al.* found an overall advantage for natural speech. However, in agreement with Luce's results, they also found that the advantage of natural speech over synthetic speech was smaller for the individual words than for the propositions. The fact that memory for words was not actually better for synthetic speech compared to natural speech in their study may reflect characteristics of the particular synthesizer used. Ralston *et al.* used Votrax, a synthesizer for which identification accuracy is fairly low, in contrast to Luce, who used MITalk, a much higher quality synthesizer. Clearly, what one wants to compare in these two studies is memory for words that were successfully identified on first encounter. The Votrax data are likely to include words that were never identified in the first place because of poor segmental intelligibility.

Taken together, these comprehension results suggest that when listeners are presented with synthetic speech, they devote relatively more attention to surface form than for natural speech. As a result, memory for surface detail is better for synthetic speech. Although the results are suggestive, they do not allow us to determine whether this memory for surface form is a memory for the acoustic signal that is preserved longer than usual and/or whether it is a memory for the verbatim wording of a sentence apart from its acoustic form. Either type of memory could be a byproduct of the additional processing that is required when there is difficulty at the input stages. Additional perceptual processing on impoverished acoustic information may result in the maintenance of the phonetic representation for a longer time in memory. Alternatively, a mechanism that makes use of the immediate semantic and syntactic context to identify words might result in the maintenance in memory of a sentence in verbatim (but not necessarily acoustic) form.

Although there is no firm evidence for either alternative presented above, some anecdotal evidence suggests that it is worth pursuing the hypothesis that something close to the acoustic signal is preserved in memory. Subjects in our experiments frequently commented

that the synthetic voice sounded as though it had a foreign accent. One of us, for example, heard an Eastern European accent in DECTalk Paul. Furthermore, Simpson and Navarro (1984), who carried out a study using semantic differential scaling found that synthetic speech occupied the same position in multidimensional space as natural speech produced by a foreign speaker. The fact that the accent comes through to conscious awareness and is retained in memory is indirect evidence that some attentional resources are devoted to the preservation and analysis of this surface information about voice quality.

As we have indicated, the allocation of resources to surface information may reflect a response to processing difficulty. It might also reflect, however, the allocation of resources to processing that is not directed toward comprehension at all. For example, subjects might be devoting some resources to the exploration of an interesting pattern of speech. In this case, the resources are not allocated to the comprehension process. Rather they are allocated to something akin to an “aesthetic” analysis of the sound pattern.¹

Summary

There are two immediate processing outcomes of the difficulty encountered in identifying phonemes in synthetic speech. The first is the use of semantic context to facilitate identification of words. The second is additional perceptual analysis on the acoustic input. Both of these processes can be expected to make use of limited processing resources. In addition, resources may be allocated to an aesthetic analysis of interesting acoustic input. In the next section, we examine the question of whether there is a resulting cost to the overall comprehension process.

The Cost of Processing Synthetic Speech

Earlier, we saw a processing cost to identifying isolated words in synthetic speech. Even synthetic words that are perfectly identifiable require more processing time for recognition. The comprehension system can partially compensate for bottom-up processing difficulties through the use of context. It is also clear from the gating experiment, however, that context does not override the need to hear more of a synthetic signal compared to a natural signal. In this section, we focus on the issue of how these processing difficulties at the phoneme and word level cause problems for later comprehension processes. We will argue that the comprehension of synthetic speech requires more processing resources than natural speech. We review two kinds of studies: Experiments that have measured processing time directly and experiments that have measured processing load per unit time.

The time course of comprehension processes

In this section, we review studies that have examined the time course of spoken language comprehension. Two tasks are discussed: a sentence verification task and a sentence-by-sentence listening time task. In the former, the subject hears a sentence and presses a response key indicating a true or false judgment. In the latter, the subject controls the rate of presentation of successive sentences. For both tasks, response time is measured from the *end* of the sentence to the button-press response. Differences in response time reflect differences in the amount of processing that took place after the sentence was presented.

We will discuss the results of these studies within the following framework. For each sentence, a subject must carry out two kinds of processes: Word-related processes that are tied to individual words and higher-level integrative processes that produce and evaluate a representation that is used to derive the full meaning of the sentence. The total amount of

¹We thank Sheri Hunnicutt for this observation.

time and resources devoted to word-related processes for a given sentence is assumed to vary directly with the number of words in the sentence. The total time and resources devoted to higher level processes are not assumed to vary directly with the number of words in the sentence (at least within fairly short sentences). We assume further that higher-level processes are initiated immediately as each word is recognized. Final sentence wrap-up processes (Just and Carpenter, 1980) and processes that assess the truth value of the sentence as a whole, however, cannot be initiated until the end of the sentence. For natural speech, these are the processes most likely to require time beyond a sentence's offset point.

Manous, Pisoni, Dedina, and Nusbaum (1985) used the sentence verification task to examine the response time pattern for natural speech compared with several different types of synthetic speech, including high and low quality synthesizers. Subjects heard a sentence, immediately pressed the true or false key, and then transcribed the sentence. Response times for incorrectly transcribed sentences were excluded from the final data analysis. Response time means are presented in Figure 8. Within the response times, an overall effect of voice was observed. Response times for natural speech were consistently shorter than any of the synthetic voices, and response times for the better quality synthetic speech were shorter than response times for the less intelligible ones. Note that differences in response times were observed in the absence of differences in transcription accuracy.

In all of the studies of processing time that we discuss, the effect of voice will be the major finding to be explained. Initially, one might hypothesize that the need for additional time after the end of the sentence is a direct consequence of the lengthened word recognition processes described earlier. That is, each synthetically produced word required additional processing or consolidation time beyond its utterance time for correct recognition to take place. The simplest form of this hypothesis, however, would predict an effect of sentence length in the data for synthetic speech. Indeed, such an effect was found within the False data for the lowest quality synthetic voices. The absence of a length effect within the other conditions, however, can be accounted for given the results of the gating studies described earlier which indicate that words can be recognized before their last sounds have been analyzed. Frequently, for natural speech, word recognition processes are completed within a word's utterance time (but see Bard, Shillcock, and Altmann, 1988; Grosjean, 1985). This means that word recognition times can be lengthened to some extent without exceeding the total utterance time for a sentence. If this is the case, then up to some limit, a slowdown in word recognition processes resulting from difficulties in recognizing synthetic words would simply use up more of a word's utterance time, leaving less time for higher level processes. When word recognition uses most of the resources available during a sentence's utterance time, then higher level processes must be carried out after the end of the sentence. This may be what is producing the voice effect in this experiment.

In a follow-up study, Pisoni, Manous, and Dedina (1987) again used the same task; this time only the highest intelligibility synthetic speech was used (DECTalk Paul), and sentence predictability was added as an additional factor. In addition, natural and synthetic sentences were equated for transcription accuracy. Response time means are presented in Figure 9. Pisoni *et al.* found that verification times were again longer for synthetic sentences compared to natural. Furthermore, for both voices, predictable sentences were faster than unpredictable ones and the two factors did not interact. The absence of an interaction was interpreted as indicating that voice and predictability affected different stages of processing. Voice was assumed to affect the word recognition stage as described above; predictability, in contrast, affects a later verification stage.

Converging evidence that synthetic speech requires more processing time is provided by recent findings of Ralston *et al.* (1991) using the sentence by sentence listening time task.

This task was intended to mirror the sentence-by-sentence reading tasks that have been used in the reading literature (e.g., Graesser, Hoffman, and Clark, 1980; Haberlandt, 1984). Subjects heard passages of connected speech presented in a natural voice or in a lower quality synthetic voice (Votrax); the passages varied in difficulty (fourth grade *vs.* college level). Subjects controlled the presentation of each successive sentence by pressing a response button. Listening times were measured from the end of each sentence to the subject's button press. Mean listening times are presented in Figure 10. Ralston *et al.* (1991) found that listening times were longer for synthetic speech and for the more difficult texts. The voice and difficulty factors did not interact. This is consistent with an analysis that places the voice effect at the word level and the difficulty effect at a level that involves integrating and inferencing about sentences.

Finally a study by Hersh and Tartaglia (1983) is relevant here, although a different kind of measure was used to assess comprehension. They varied speaking rate for a set of meaningful texts and examined comprehension performance on a multiple-choice comprehension test administered after each text. Stimuli were presented only in high quality synthetic voices (DECTalk Paul and Betty), and patterns of performance were compared to patterns collected earlier on compressed natural speech by Fairbanks, Guttman, and Miron (1957). Fairbanks *et al.* found that comprehension performance for natural speech remained fairly constant at increased speaking rates until the rate exceeded 300 wpm. In contrast, Hersh and Tartaglia found that comprehension performance for synthetic speech, even very high quality synthetic speech, remained constant at 150 and 180 wpm but then dropped off fairly dramatically well before 300 wpm. Although different stimuli were used in the two experiments, comprehension performance at 150 and 180 wpm was comparable for both sets of stimuli (about 65% accuracy).

Hersh and Tartaglia's finding that performance on synthetic speech remained constant at 150 and 180 wpm suggests that, at slow speeds, the comprehension system does not need the total utterance time to complete all of the necessary processing for a sentence produced by a high-quality speech synthesizer. Their finding that performance for synthetic speech begins to degrade at much slower presentation rates than natural speech suggests that the relative amount of unneeded time is much less for synthetic speech than natural speech, converging nicely with the listening time studies described earlier.

In summary, because the listener does not normally have control over the rate of presentation of spoken language, the critical resource during spoken language comprehension is time. Within the time it takes the speaker to utter a sentence, the listener must not only recognize the words but also carry out some set of higher level syntactic, semantic, and inferential processes to assign a meaning to the input signal. The point of this section has been to suggest that because low-level word recognition processes for synthetic speech require more processing time, there is less time left over for higher-level processes during a sentence's utterance time. As a result, more processes must await completion after the sentence has been produced.

Processing capacity

Additional evidence that synthetic speech requires more processing resources can be found in studies using a secondary task methodology. In these experiments, listeners must comprehend a text presented in a natural or synthetic voice while also monitoring for a particular auditory stimulus. In addition to the sentence by sentence listening time task described earlier, Ralston *et al.* (1991) also included conditions in which subjects had to monitor passages for particular target words (either two or four words at a time) while simultaneously comprehending the text. Subjects were instructed to press a response button whenever they heard one of the target words in the text. Response time and accuracy were

recorded. In these conditions, a memory pre-load was combined with a secondary task that was carried out concurrently with comprehension processes.

Ralston *et al.* found that listeners were faster and more accurate in responding to the target words in natural speech compared to synthetic speech. They were also faster in monitoring for two words compared to four. The main effect of voice may simply reflect the longer time it takes to recognize a synthetic target word compared to a natural target word. In addition, the reaction time data yielded an interaction of voice and text difficulty (fourth grade *vs.* college) which resulted from the slow response times for the college level texts presented using the synthetic voice. Two accounts of this interaction are possible. First, it may simply reflect difficulty in identifying unintelligible words in Votrax. Alternatively, the interaction may reflect several resource-demanding processes that are carried out simultaneously and which compete for limited resources: the process of identifying synthetically produced words, the process of comprehending a difficult text, and the process of monitoring for the presentation of a particular word (the secondary task).

Ralston, Lively, Pisoni, and Rivera (1990) conducted another monitoring experiment to dissociate the alternatives discussed above. In this experiment, subjects heard the same set of passages used in the earlier experiment, but now they monitored for a non-speech sound (a click) rather than for a word. Thus, in all conditions (natural as well as synthetic) the target items for the secondary task were identical. Any differences across experimental conditions in responses on the secondary task cannot be due to differences in the ease of identifying the target stimuli. Ralston *et al.* found that subjects were slower to respond to clicks in the synthetic voice condition than the natural condition. This finding suggests that more resources are used during a synthetic sentence's utterance time, leaving fewer resources for carrying out the secondary task.

So far, the claims about the role of processing capacity have been examined using tasks in which extra capacity is needed to carry out some secondary task that is dissociated from the realm of comprehension. One can, however, think of the whole language comprehension process as a multi-task process in which various subprocesses are competing for limited resources (Perfetti, 1985; Stanovich, 1980). To date, we know of no studies that have directly addressed this question by targeting a particular resource-demanding comprehension process and looking at its performance when text is presented in a synthetic *vs.* natural voice.

Post-perceptual measures of comprehension

We have developed an account which claims that synthetic speech requires additional processing time beyond utterance time for comprehension. When sentences are heard in a multi-sentence context where the listener does not control presentation time, the system should become overloaded. Some necessary processes will simply not get completed in time. Given this emerging account of comprehension processes for synthetic speech, one would expect to find evidence that the final output of comprehension of a whole discourse of synthetic speech would be impoverished in some way in comparison to natural speech. That is, if one simply asked subjects to listen to a passage of connected text (where subjects do not control sentence-by-sentence presentation time) and then answer some questions about it, evidence of comprehension difficulties for synthetic speech should emerge. Here, however, the pattern of results reported in the literature is surprisingly mixed. Frequently, one finds little difference in performance for synthetic speech compared with natural speech. For example, Nye *et al.* (1975), in their evaluation of the Haskins synthesizer, asked subjects to listen to texts taken from college-level reading tests and then had them answer multiple-choice questions. Subjects were allowed to replay the texts while answering the questions. Nye *et al.* found no difference in accuracy for synthetic speech compared with natural

speech. This initial study was followed by several studies in which negligible differences were also found (McHugh, 1976; Ingemann, 1978; Jenkins and Franklin, 1982; Pisoni and Hunnicutt, 1980).

When differences were found, they disappeared after minimal practice. For example, McHugh (1976), using fourth grade texts as stimulus materials, found an effect of voice on her two practice paragraphs that disappeared for the main experimental paragraphs. Jenkins and Franklin (1982), using a text taken from the children's section of a newspaper, found no difference in free recall for synthetic speech compared with natural speech when the subjects hearing the synthetic speech had heard 20 synthetic sentences one week before testing. Thus, with minimal practice, comprehension performance for synthetic and natural speech appear to be roughly equivalent when accuracy measures are used.

One might raise several methodological criticisms about the studies in which no differences were found. In only one of the above studies (e.g., Nye *et al.*, 1975) is there any mention of pretesting the questions to make sure that subjects could not answer them without having heard the text. In some cases, the text may have been on such a highly familiar topic that subjects could guess at the answers to the comprehension questions. In many cases, the subjects may have been able to reconstruct what must have been in the passage based on fragmented memory of the text. In two of the studies (Ingemann, 1978; Nye *et al.*, 1975) subjects were tested after hearing the text twice. This may have eliminated any differences that existed on initial presentation.

In contrast to these findings, a number of more recent comprehension studies have reported lower performance on comprehension tests for synthetic speech compared to natural. Ralston *et al.* (1991) found that recognition memory for propositions and for individual words was overall better for natural speech compared with Votrax speech; this was true for easy texts (fourth grade level) and for college level materials. Questions used in the Ralston *et al.* study were pre-tested to ensure that subjects could not answer them without actually listening to the passages. The Ralston *et al.* findings could be due to a lower identification accuracy rate for Votrax. But other studies using higher quality synthesizers have also found a similar disadvantage for synthetic speech in tests of comprehension and recognition memory. Luce (1981) found that memory for meaning was better for natural speech than for MITalk. Moody and Joost (1986) found an advantage for natural speech compared to DECTalk for questions about information that was stated explicitly in the passages. Surprisingly, they found no difference between the two voices for more difficult inferential questions, but this result may simply be due to a floor effect.

In summary, the account that we have developed so far leads us to predict that the final product of the comprehension process should be more impoverished, fragmented, and less robust for passages of synthetic speech than for natural speech. This fragmentation should occur because even when words are identifiable, more processing time is spent on word recognition, leaving less time for the integration and inferencing necessary for normal comprehension (Myers and Duffy, 1990) and retention of what was comprehended. The fact that synthetic speech is not always at a disadvantage on these post-perceptual tests is likely to reflect an insensitivity in the comprehension test and response measures used. In those studies in which researchers have been very careful to select texts of appropriate difficulty and to pre-test their comprehension questions, differences in comprehension between natural and synthetic speech have been observed not only for low-quality synthesizers like Votrax (Ralston *et al.*, 1991) but also for a high quality synthesizer like MITalk (Luce, 1981). We would expect to see these differences if researchers begin to probe explicitly for the products of resource-demanding comprehension processes that require either the integration of

information across several sentences or the drawing of inferences based on prior knowledge in long-term memory.

Practice Effects

In the research reviewed above, we have focused primarily on documenting the immediate effect of synthetic speech input on the language comprehension system. In this section, we consider what happens to this system over a longer time period after it has become practiced at listening to synthetic speech. Three hypotheses are worth considering in this regard. The first and least likely is that the system changes very little because the synthetic input is so impoverished that there is little room for improvement in processing. Second, the system may become skilled at making use of compensatory strategies: using semantic context for word recognition and holding a surface form of the input in memory to allow prolonged perceptual analysis. Third, the system may become skilled at recognizing synthetic phonemes, abandoning dependency on semantic context for word recognition and coming to resemble more closely the highly efficient system already developed for the processing of natural speech.

Effects of practice and familiarity on accuracy measures

A number of the studies already reviewed provide, not surprisingly, evidence against the first hypothesis above, namely, that the perceptual system changes little with practice. Again and again, in experiments that provide minimal experience with synthetic speech, performance was found to improve dramatically even after a short period of time. Nye and Gaitenby (1973) found significant improvement across trials on the MRT for synthetic speech but no change for natural speech. They also found that practiced subjects made fewer errors than unpracticed subjects on synthetic sentences that were semantically anomalous (Nye and Gaitenby, 1974). The largest differences were found for a subset of consonants and vowels that clearly were exhibiting learning effects.

Pisoni and Hunnicutt (1980) found a small but reliable improvement in performance for both semantically anomalous and semantically normal sentences presented using MITalk when performance on the second half of a 100-sentence transcription task was compared with performance on the first half (see also Pisoni, 1987). Pierce and Remington (1984) and Simpson (1975) also found that subjects improved from the first to second trial in transcribing air traffic control messages. Furthermore, Simpson found that the advantage of expertise in transcribing air traffic control messages in synthetic speech was smaller on the second presentations of these messages.

Several comprehension studies have also observed practice effects. Recall that McHugh (1976) found no effect of voice on accuracy in answering short answer questions except for her two practice paragraphs. This suggests that by the time subjects began to listen to the experimental paragraphs, their ability to process synthetic speech had already improved greatly. Jenkins and Franklin (1982) found that minimal practice (hearing a list of 20 words in synthetic speech one week before the experiment) resulted in equivalent performance in recall measures for listeners hearing natural speech compared to listeners hearing synthetic speech. Finally, Pisoni and Hunnicutt (1980) examined performance on multiple choice comprehension questions for the first half of their texts compared with the second half. They found significant improvement in performance for subjects who listened to synthetic speech but no improvement for subjects who listened to natural speech.

What are subjects learning about synthetic speech? Some of the studies provide evidence that subjects are gaining skill in recognizing specific phonemes. For example, Nye and Gaitenby (1974) found that practiced subjects performed better at recognizing five of the

consonants (j, m, n, v, w) and four vowel sounds. Since they were using semantically anomalous sentences, this improvement cannot be due to some strategic use of semantic context. Rather, it must be that subjects learned something quite specific about the mapping between the synthetic speech sounds and the intended phonemes.

Rosson (1985) also found perceptual learning that was specific to certain phonemes. She trained subjects using a 20-minute sentence transcription task. During training, subjects heard only half of the complete set of “sounds” (Rosson’s term) that were included among the final test sentences; training sentences were presented in either natural or synthetic voice. The final test was a transcription task with sentences presented only in the synthetic voice. Not surprisingly, subjects trained on the synthetic voice did better than those who had heard the same training sentences in a natural voice. Subjects also performed better transcribing the sounds presented during training than the untrained sounds. But the results were complicated by a higher-level interaction involving materials sets. Further analyses indicated that for subjects trained on the synthetic voice, training only facilitated the identification of some of the sounds. That is, some sounds showed no effect of training, either because they were already highly intelligible or because they never improved. Rosson suggested that within the set of synthetic speech sounds there are “learnable” sounds that will benefit from training. Of course, an interesting question remains: Of those sounds that are fairly unintelligible initially, what makes some learnable within a short period (Rosson’s 20 minutes of practice) and others unlearnable within that same time?

Long-term training studies

Schwab, Nusbaum, and Pisoni (1985) studied perceptual learning of Votrax synthetic speech over an eight day period. On Day 1 (Pre-test) and Day 10 (Post-test), subjects’ performance was measured on a set of tasks involving synthetic speech. Three groups of subjects were run. One group was explicitly trained on Days 2 through 9 with Votrax speech. A second group received the same training but with natural speech. A third group received no training at all. Performance on Day 10 compared with Day 1 was assessed for each group. The natural speech training group was included to measure the extent to which improvement simply reflected practice with the specific tasks rather than with the synthetic speech. Five tasks were included in the tests: (1) intelligibility test of phonetically-balanced words, (2) the MRT, (3) transcription of semantically well-formed sentences, (4) transcription of semantically anomalous sentences, and (5) answering yes/no comprehension questions following prose passages. All tasks except the MRT task were also used during the eight days of training, although with different stimulus materials on each day. Accuracy was the dependent measure for all tasks; in addition, for the MRT and the comprehension tests, a response time measure was collected.

In the transcription and intelligibility tests (Tasks 1–4 above), the greatest improvement in accuracy from Day 1 to Day 10 was for the group trained specifically with Votrax synthetic speech. Unexpectedly, for the yes/no comprehension question task (Task 5 above), all groups performed better on Day 1 than Day 10. Taken together, the results indicate that although there is evidence of some learning of the task itself, as indicated by a modest improvement in performance for the other two groups on some of the tasks, there is also perceptual learning that is quite specific to synthetic speech.

These learning data can be used to evaluate the hypotheses outlined at the beginning of this section. Of special interest is the improvement in performance on the semantically meaningful sentences compared with the anomalous sentences. If the practice effects are reflecting better use of semantic context, then one would expect to find more improvement for meaningful sentences where context can be used compared with anomalous sentences where semantic context is useless. The data do not support this prediction. While the

Votrax-trained group displayed a 35.8% improvement from Day 1 to Day 10 in transcription accuracy for the meaningful sentences, they also showed a 50.2% improvement for the anomalous sentences. In fact, on Day 1 they were better at transcribing the meaningful sentences (42% accuracy) compared to the anomalous (27% accuracy); but by Day 10 there was no difference in performance on the two sentence types (77.8% vs. 76.9%). This lack of difference between the meaningful and anomalous sentences on Day 10 is itself further evidence for the decreased dependence upon context as a function of training.

Under the hypothesis that subjects get better at using context, one would also expect little improvement for words presented in isolation. Again, the data provide no support for such a claim. The Votrax-trained subjects showed a 40% improvement in accuracy at identifying the phonetically-balanced words in isolation on Day 10, compared with a 5% improvement for the other two groups. The Votrax-trained subjects also exhibited a 16.9% increase in MRT accuracy compared with no improvement for the other two groups. This overall pattern is consistent with the hypothesis that, with practice, listeners make less use of semantic context for perceiving synthetic speech and make more effective use of information in the acoustic signal itself. Subjects learned the specific pairing of acoustic cues to phonemes that is idiosyncratic to Votrax synthetic speech.

Examination of the response time data provides evidence that, in addition to becoming more accurate, listeners also become faster at perceiving synthetic speech. In the MRT task, time to press the button corresponding to the correct response alternative was recorded. The Votrax-trained group was 462 msec faster in responding on Day 10 compared with Day 1. The natural voice-trained group was also faster on Day 10 than Day 1, but this difference was nonsignificant. The reduction in response time for the Votrax-trained group cannot be due to more efficient use of semantic context since there is none available in this task; the words are presented in isolation. The decrease in response latency also cannot be due to practice in choosing one out of six response alternatives because the natural speech group showed only minimal improvement with the same training procedures. Response times to the yes/no comprehension questions revealed no systematic patterns. Thus, the response time difference on the MRT task seems to be the only evidence available to support the claim that with practice listeners become faster at analyzing the acoustic-phonetic structure of synthetic speech. This finding is an important one, however, because it suggests that the result of practice is a reduction in the amount of time that must be devoted to recognizing words. Thus, more of the utterance time for synthetic speech can be devoted to higher level comprehension processes that involve integration and inference.

Further evidence that listeners become more proficient at analyzing the acoustic-phonetic structure of synthetic speech with practice comes from a study by Greenspan, Nusbaum, and Pisoni (1988). In their experiments, subjects received four days of training listening to Votrax speech. In the pre- and post-tests (on Days 1 and 6), subjects were asked to transcribe semantically well-formed and anomalous sentences, a list of phonetically balanced words, and to carry out the MRT task. Two training factors were varied orthogonally. First, training stimuli were either isolated words or sentences. Second, training stimuli were either repeated or not repeated across days of training. Thus, both stimulus type and stimulus variability were varied in this design.

In the post-test, all of the trained subjects improved compared to a group of untrained controls who were tested on Days 1 and 6 with no intervening training. Furthermore, for the post-tests on isolated words, all trained subjects performed equally well. In contrast, for the post-test sentence transcription tasks, the results were more complex. For the semantically well-formed sentences, only the sentence-trained groups improved. For the anomalous sentences, all groups improved but the sentence-trained groups improved the most. The

amount of improvement for the two word-trained groups did not differ from that of the untrained control group.

This pattern of results suggests that both word and sentence training facilitated isolated word recognition. In contrast, only sentence training facilitated sentence transcription. As in the previous study by Schwab *et al.* (1985), the improvement in word recognition with practice suggests that subjects are not merely learning how to use semantic context better to aid them in word recognition. Improvement in recognition of isolated words can only occur if subjects are learning something about the mapping between the acoustic-phonetic cues present in the synthetic speech and the intended phonemes. The results for the sentence transcription task suggest that subjects are also learning to segment the acoustic-phonetic stream into individual words. Improvement in segmentation only comes with practice on sentences. The fact that segmentation ability improved with the anomalous as well as the semantically meaningful sentences again supports the claim that subjects are learning very detailed information about the acoustic-phonetic structure of the synthetic voice and are not relying increasingly on semantic context.

Since variability in the training stimuli had no effect in this experiment, Greenspan *et al.* (1988) reduced the repeated training set even further in a second experiment. Here the authors did find that subjects trained on a small, repeated set of stimuli improved less than subjects trained on stimuli that were not repeated during training. Greenspan *et al.* argue that this finding provides evidence that subjects are acquiring information about the acoustic-phonetic structure of the synthetic voice and that subjects need a fairly large sample of different utterances in this voice to extract the relevant structure.

Taken together, the results from the training studies suggest that listeners become much more proficient at analyzing the acoustic-phonetic input of a synthetic voice with practice. As a result, the LCS moves toward a mode of operation more typical of that used to process natural speech. We should note that, under this interpretation, proficiency with one synthesis system should not necessarily generalize to other systems. During training, the listener is acquiring knowledge that is highly specific to the acoustic-phonetic structure of a particular rule system used to generate synthetic speech. Thus, there should be no evidence of transfer of this expertise and knowledge to another synthetic voice that uses a different rule system. To our knowledge, no studies have been conducted to assess this prediction.

Summary and Conclusions

Summary

In this paper, we have reviewed a large number of studies in which the perception and comprehension of natural and synthetic speech have been compared. The goal of our review has been to understand how synthetic speech affects the operation of the LCS. We have developed the following general account. When synthetic speech is presented as input to the LCS, the phonetic processor is immediately confronted with difficulty in identifying phonemes from the acoustic stream; as a consequence, the word recognition processor has difficulty extracting the intended words. This difficulty results in an increase in time needed for the analysis of the acoustic-phonetic information. In some cases, this difficulty results in an inability to identify a word based on the acoustic-phonetic input; in this situation, the LCS compensates by making use of semantic context to identify the word. The result of these difficulties is a delay in word recognition, so that a synthetically produced word is identified at a later point in the utterance than a naturally produced word.

This processing delay in word recognition results in the allocation of more resources to word recognition. Within the duration of a given utterance, less time is therefore left for higher

level comprehension processes. When utterance time is not sufficient for carrying out all of the necessary comprehension processes, comprehension performance suffers and the resulting memory representation for the discourse is fragmented and incomplete.

With training, the phonetic processor learns to pair particular acoustic cues with intended phonemes for a given synthetic voice. As a result, words can be identified more accurately and more rapidly. The use of context becomes less important now, and the LCS moves toward the more typical data-driven state in which it normally operates when analyzing naturally produced speech.

Other modes of speech

This review has focused on the processing of synthetic speech produced by rule. We do not mean to suggest, however, that synthetic speech poses entirely unique problems for the LCS. There are, of course, other forms of speech that share the critical features of synthetic speech that affect processing: poor segmental quality and, in many cases, poor prosody. Such forms of speech include the speech of deaf speakers, non-native speakers of a language, speakers with speech impediments, young children, etc. In all cases, these other forms of speech pose difficulties at the phoneme and word recognition level. The existing literature on the processing of these kinds of speech includes effects that mirror those for synthetic speech. For example, intelligibility is affected by context. Gladstein, Membrino, and Flege (1983, cited in Flege, 1988) found that English words spoken by native Vietnamese speakers were more intelligible in predictable contexts than in unpredictable contexts. McGarr (1981) and Monsen (1983) found similar context effects on the intelligibility of sentences produced by deaf speakers. They also observed that listeners who were practiced at listening to deaf speech found the speech more intelligible. Furthermore, as with synthetic speech, McGarr found that practice raises the intelligibility of unpredictable sentences as much as predictable. Converging with the conclusions from the synthetic speech literature, McGarr concludes that experience is producing a more skilled perceptual analysis of the input rather than a more effective use of context.

Another listening situation likely to produce processing difficulties that might be similar to those produced by synthetic speech is that of a language learner hearing speech in the language that is being learned (L2). Here, phoneme and word recognition processes should be delayed (see review by Flege, 1988) and should be sensitive to context. Kilborn (1992), however, found that the specific pattern of delay is not a simple one. He investigated the effect of context on English word monitoring performance for German-English bilinguals compared to English monolinguals. He found that response times were significantly longer for the bilinguals only when semantically meaningful sentences were used; no differences between the two groups emerged when random word strings or semantically anomalous but syntactically well-formed sentences were used. On the surface, this seems at odds with the context effects observed in the literature on intelligibility of synthetic speech where the addition of meaningful context always has a greater effect for the more difficult form of speech. Kilborn observes, however, that speech in L2 poses additional difficulties for the LCS (difficulties that are not created by synthetic speech) because the LCS may impose inappropriate syntactic assumptions derived from L1 on the input (see also Kilborn, 1989). Thus the process of integrating syntax and meaning that is required to make effective use of a meaningful context is delayed beyond the point of word recognition.

Finally, some intriguing contrasts in intelligibility have been observed in a gating study employing words produced by native English speakers with no speech or hearing impairment (McAllister, 1991). McAllister found that subjects were more successful in identifying words excised from read speech than from spontaneous speech. Within these two speech modes, subjects were more successful identifying words whose initial syllables were

stressed compared with unstressed. Adding meaningful context improved performance in general, although the improvement was smallest for the initially stressed words in the read speech. In some ways these results parallel those from the gating study discussed earlier in which natural and synthetic speech was compared.

In summary, there are other modes of speech which cause difficulties for the LCS at the phoneme and word recognition level. The critical question is to what extent research on these speech modes converges with that in the synthetic speech processing literature. Given the analysis we have developed, we would expect that any mode that caused difficulties in low-level processing would have a similar impact on higher level comprehension processes. Thus, it would be useful to extend the present analysis to these other kinds of speech in order to determine whether converging evidence exists for the general claim that low level processing difficulties have implications for higher level comprehension processes.

Future directions for research

Although, in its general form, our account of synthetic speech comprehension is supported by the research to date, there are a number of assumptions and predictions that have not been directly tested. In our account, we have suggested that difficulties at low levels of processing such as phonetic perception and word recognition have an effect on later comprehension processes by depriving them of resources, in particular needed processing time. This theoretical account rests on some assumptions that we outlined about the time course and scheduling of these various processes in the course of language comprehension. To date, however, there has actually been very little research on the time course and scheduling of comprehension processes for natural speech. Measures using sentence verification time or sentence-by-sentence listening time are relatively new techniques, and it is not clear exactly what is being measured in these tasks. At the present time, we have no systematic studies of the effects of potentially critical factors such as sentence length, syntactic complexity, and inferential complexity on these measures. The development of a more detailed model of real-time comprehension processes for natural speech will make it much easier to determine how those processes are affected when phoneme and word identification processes are delayed because of degraded or impoverished input.

For research that is directly focused on the LCS's response to synthetic speech input, we see four issues as promising directions for future research:

Targeted higher level processes—If a delay in word recognition deprives higher level processes of needed time resources, then it should be possible to target certain resource-demanding comprehension processes and provide evidence (by probing for specific products of comprehension) that particular comprehension processes were either never carried out, or were executed only incompletely. For the post-perceptual comprehension measures discussed earlier, investigators did not test for the products of well-specified comprehension processes. For example, was the correct antecedent retrieved for a particular noun or pronoun? Was a specific inference drawn that was required for comprehension of a particular sentence in the text? Was an earlier proposition retrieved in order to make a causal connection? Given the account of processing that we have developed, we would predict that one would find more evidence of the failures of such resource-demanding comprehension processes when the input is synthetic speech, particularly poor-quality synthetic speech.

Effects of prosody on comprehension—We have built our account of synthetic speech comprehension on difficulties caused by delayed word recognition because of the extensive focus in the literature on the intelligibility of phonemes and words in synthetic speech. Another critical difference, however, that distinguishes synthetic speech and natural

speech is the absence of appropriate prosody. This problem is just now receiving extensive attention in the development of synthetic speech systems (e.g., Auberge, 1992; Hirschberg, 1992; Monaghan, 1992; Padeloup, 1992). One common problem within the synthetic speech systems is the failure to use contrastive stress and to de-accent given information (Monaghan and Ladd, 1990). This is a difficult problem to solve because determining appropriate stress patterns frequently requires that the system “understands” what it is saying (Bolinger, 1972), although there are efforts underway to produce a set of heuristics that would allow a synthetic speech system to identify the given information in many cases without understanding (Hirschberg, 1992; Monaghan, 1992).

Research on the effect of the prosody of synthetic speech on the LCS indicates that prosody affects both rated “attractiveness” of synthesized speech (Terken and Lemeer, 1988) and intelligibility (Benoit, 1990). Furthermore, in both of these studies, prosody interacted with segmental quality. Specifically, the introduction of natural prosody has little effect on intelligibility or rated attractiveness of speech with fairly low segmental quality, but it does increase the intelligibility or rated attractiveness of higher quality speech. Slowiaczek and Nusbaum (1985) found that pitch contour (normal vs. monotone) had no effect on intelligibility of syntactically simple sentences but did affect intelligibility of syntactically complex sentences. Larkey and Danly (1983), using vocoded speech, found longer sentence verification times for sentences presented in a monotone compared with normal prosody.

Research on prosody in natural speech also provides evidence that prosody affects the LCS. For example, Tyler and Warren (1987) found that word monitoring times were longer when the overall prosody of the sentence was disrupted and when just the local prosodic unit in which the target word appeared was disrupted. A number of studies have found that prosody affects response time in a phoneme monitoring task (e.g., Cutler, 1976; Mehta and Cutler, 1988; Shields, McHugh, and Martin, 1974). Furthermore, prosodic factors (e.g., word duration and pitch) are used as a cue to the given/new structure within a sentence (Fowler and Housum, 1987; Terken and Nootboom, 1987).

All of these studies suggest that the lack of appropriate prosody must contribute significantly to the comprehension difficulties observed for synthetic speech. The absence of appropriate prosodic cues may result in a lengthening of syntactic parsing processes and integrating processes as well as a lengthening of word recognition processes. This means that the higher-level processes of the LCS have additional resource needs as well as the lower-level processes. As synthetic speech systems become more successful in producing utterances with appropriate prosody, we would expect to observe reductions in such behavioral measures as sentence-by-sentence listening times, sentence verification times, and word monitoring times.

Effects of training—Although the previous training studies suggest that listeners become better at analyzing the acoustic-phonetic structure of a synthetic voice, we know very little about how training actually affects comprehension processes. As proficiency in analyzing the acoustic-phonetic structure of synthetic speech develops, we would expect a decrease in the amount of resources required for comprehension of synthetic speech. Thus, for example, we would expect response times for a secondary task like click monitoring to be faster after training, listening times to be shorter, and comprehension scores to be higher.

Ease of processing—If synthetic speech makes more resource demands than natural speech, then we should find evidence for this in various measures of mental workload or cognitive effort. The secondary task measures described earlier provide some evidence for increased resource use as do the listening time measures. But the greater demands on

processing resources should also produce greater fatigue effects in the long run that might be detected by other cognitive measures. To our knowledge no one has yet addressed this issue.

Conclusion

In conclusion, there is much research yet to be done in this area. The present paper has attempted to provide a theoretical framework for conceptualizing comprehension of synthetic speech, a framework that may be useful in guiding future studies. We feel that research on comprehension of synthetic speech and comparisons with natural speech will contribute to our understanding of the language comprehension system as a whole and, in particular, to our understanding of how comprehension processes operate in real time with limited resources. The bulk of research on the perception of synthetic speech over the last decade has been concerned primarily with measuring segmental intelligibility of isolated words. In comparison, very little research has been carried out on comprehension and memory for sentences and long passages of connected meaningful text. We hope that researchers will turn their attention to this important problem and to the accompanying task of developing new experimental techniques and methodologies that can be used to assess the speech comprehension process as it occurs in real-time.

References

- Allen, J.; Hunnicutt, MS.; Klatt, D. From Text to Speech: The MITalk System. Cambridge, UK: Cambridge University Press; 1987.
- Altmann, GTM., editor. Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives. Cambridge, MA: MIT Press; 1990.
- Auberge, V. Developing a structured lexicon for synthesis of prosody. In: Bailly, G.; Benoit, C.; Sawallis, TR., editors. Talking Machines: Theories, Models, and Designs. Amsterdam: North-Holland; 1992. p. 307-321.
- Bard EG, Shillcock RC, Altmann GTM. The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*. 1988; 44:395–408. [PubMed: 3226889]
- Balota, DA.; Flores D'Arcais, G.; Rayner, K., editors. Comprehension Processes in Reading. Hillsdale, NJ: Erlbaum; 1990.
- Benoit C. An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. *Speech Communication*. 1990; 9:293–304.
- Bolinger DL. Accent is predictable (if you're a mindreader). *Language*. 1972; 48:633–644.
- Clark JF, Dermody P, Palethorpe S. Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons. *Journal of the Acoustical Society of America*. 1985; 78:458–462.
- Coltheart, M., editor. Attention and Performance XII: The Psychology of Reading. Hove, UK: Erlbaum; 1987.
- Cooper FS, Delattre PC, Liberman AM, Borst JM, Gerstman LJ. Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*. 1952; 24:597–606.
- Cooper FS, Liberman AM, Borst JM. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences*. 1951; 37:318–325.
- Cutler A. Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*. 1976; 20:55–60.
- Duffy SA, Henderson JM, Morris RK. Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1989; 15:791–801.
- Duffy, SA.; Pisoni, DB. Research on Speech Perception Progress Report No. 17. Bloomington, IN: Indiana University; 1991. Effects of sentence context on the signal duration required to identify natural and synthetic words; p. 341-354.

- Duffy SA, Rayner K. Eye movements and anaphor resolution: Effects of antecedent typicality and distance. *Language and Speech*. 1990; 33:103–119. [PubMed: 2283922]
- Fairbanks B, Guttman N, Miron MS. The effects of time compression upon the comprehension of connected speech. *Journal of Speech and Hearing Disorders*. 1957; 22:10–19. [PubMed: 13406818]
- Flege, J. The production and perception of foreign language speech sounds. In: Winitz, H., editor. *Human Communication and Its Disorders. A Review 1988*. Norwood, NJ: Ablex; 1988.
- Fodor, JA. *The Modularity of Mind*. Cambridge, MA: MIT Press; 1983.
- Forster, KI. Levels of processing and the structure of the language processor. In: Cooper, WE.; Walker, E., editors. *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Hillsdale, NJ: Erlbaum; 1979. p. 27–85.
- Fowler CA, Housum J. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and language*. 1987; 26:489–504.
- Frauenfelder, UH.; Tyler, LK. *Spoken Word Recognition*. Cambridge, MA: MIT Press; 1987.
- Ganong WF. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*. 1980; 6:110–125. [PubMed: 6444985]
- Gernsbacher MA, Faust ME. The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1991; 17:245–262.
- Gernsbacher MA, Varner KR, Faust ME. Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1990; 16:430–445.
- Gladstein, L.; Membrino, M.; Flege, J. The Effect of Context on Americans' Perceptions of Word-Final Stop Devoicing by Vietnamese L2 Learners. Washington, DC: Department of Linguistics, Georgetown University; 1983. Unpublished manuscript.
- Goldinger SD, Pisoni DB, Logan JS. On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1991; 17:152–162.
- Graesser AC, Hoffman NL, Clark LF. Structural components in reading times. *Journal of Verbal Learning and Verbal Behavior*. 1980; 19:135–151.
- Greenspan SL, Nusbaum HC, Pisoni DB. Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1988; 14:421–433.
- Grosjean F. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*. 1980; 28:267–283. [PubMed: 7465310]
- Grosjean F. The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics*. 1985; 38:299–310. [PubMed: 3831907]
- Haberlandt, K. Components of sentence and word reading times. In: Kieras, DE.; Just, MA., editors. *New Methods in Reading Comprehension Research*. Hillsdale, NJ: Erlbaum; 1984. p. 219–252.
- Hersh, HM.; Tartaglia, L. User Research Group, Corporate Research and Architecture. Maynard MA: Digital Equipment Corp.; 1983. *Understanding Synthetic Speech*.
- Hirschberg, J. Using discourse context to guide pitch accent decisions in synthetic speech. In: Bailly, G.; Benoit, C.; Sawallis, TR., editors. *Talking Machines: Theories, Models, and Designs*. Amsterdam: North-Holland; 1992. p. 367–376.
- Hoover J, Reichle J, Van Tasell D, Cole D. The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*. 1987; 30:425–431. [PubMed: 2959817]
- Ingemann, F. Haskins Laboratories Status Report on Speech Research, SR—54. New Haven, CT: Haskins Laboratories; 1978. *Speech synthesis by rule using the FOVE program*; p. 165–173.
- Jenkins JJ, Franklin LD. Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society*. 1982; 20:203–206.
- Just MA, Carpenter PA. A theory of reading: From eye fixations to comprehension. *Psychological Review*. 1980; 87:329–354. [PubMed: 7413885]
- Just, MA.; Carpenter, P-A. *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon; 1987.
- Kahneman, D. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall; 1973.

- Kilborn K. Sentence processing in a second language: The timing of transfer. *Language and Speech*. 1989; 32:1–23. [PubMed: 2622299]
- Kilborn, K. On-line integration of grammatical information in a second language. In: Harris, R.J., editor. *Cognitive Processing in Bilinguals*. Amsterdam: North-Holland; 1992. p. 337–350.
- Kintsch W, Van Dijk TA. Toward a model of text comprehension and production. *Psychological Review*. 1978; 85:363–394.
- Klatt DH. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*. 1987; 82:737–793. [PubMed: 2958525]
- Larkey, L.; Danly, M. MIT Speech Group Working Papers. Vol. Vol. II. Cambridge, MA: Research Laboratory of Electronics, MIT; 1983. Fundamental frequency and sentence comprehension; p. 25–39.
- Lieberman AM, Ingemann F, Lisker L, Delattre PC, Cooper FS. Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America*. 1959; 31:1490–1499.
- Logan JS, Greene BG, Pisoni DB. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*. 1989; 86:566–581. [PubMed: 2527884]
- Luce, PA. Research on Speech Perception Progress Report No. 7. Bloomington, IN: Indiana University; 1981. Comprehension of fluent synthetic speech produced by rule; p. 229–242.
- Luce PA, Feustel TC, Pisoni DB. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*. 1983; 25:17–32. [PubMed: 6840769]
- Luce, PA.; Pisoni, DB. Research on Speech Perception Progress Report No. 9. Bloomington, IN: Indiana University; 1983. Capacity-demanding encoding of synthetic speech in serial-ordered recall; p. 295–309.
- Manous, LM.; Pisoni, DB. Research on Speech Perception Progress Report No. 10. Bloomington, IN: Indiana University; 1984. Effects of signal duration on the perception of natural and synthetic speech; p. 311–321.
- Manous, LM.; Pisoni, DB.; Dedina, M.J.; Nusbaum, HC. Research on Speech Perception Progress Report No. 11. Bloomington, IN: Indiana University; 1985. Comprehension of natural and synthetic speech using a sentence verification task; p. 33–57.
- Marslen-Wilson, W., editor. *Lexical Representation and Process*. Cambridge, MA: MIT Press; 1989.
- Marslen-Wilson W, Welsh A. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*. 1978; 18:1–86.
- Martin CS, Mullennix JW, Pisoni DB, Summers WV. Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1989; 15:676–684.
- McAllister J. The processing of lexically stressed syllables in read and spontaneous speech. *Language and Speech*. 1991; 34:1–26.
- McGarr NS. The effect of context on the intelligibility of hearing and deaf children's speech. *Language and Speech*. 1981; 24:255–264. [PubMed: 7329194]
- McHugh, A. Listener Preference and Comprehension Tests of Stress Algorithms for a Text-to-Phonetic Speech Synthesis Program (Naval Research Laboratory Report 8015. Washington, DC: Naval Research Laboratory; 1976.
- McKoon G, Ratcliff R. The comprehension processes and memory structures involved in anaphoric reference. *Journal of Verbal Learning and Verbal Behavior*. 1980; 19:668–682.
- Mehta G, Cutler A. Detection of target phonemes in spontaneous and read speech. *Language and Speech*. 1988; 31:135–156. [PubMed: 3256770]
- Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*. 1951; 41:329–335. [PubMed: 14861384]
- Mirenda P, Beukelman DR. A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*. 1987; 3:120–128.
- Monaghan, AIC. Heuristic strategies for the higher-level analysis of unrestricted text. In: Bailly, G.; Benoit, C.; Sawallis, TR., editors. *Talking Machines: Theories, Models, and Designs*. Amsterdam: North-Holland; 1992. p. 143–161.

- Monaghan AIC, Ladd DR. Symbolic output as the basis for evaluating intonation in text-to-speech systems. *Speech Communication*. 1990; 9:305–314.
- Monsen RB. The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders*. 1983; 48:286–296. [PubMed: 6621019]
- Moody, T.; Joost, M. Synthesized speech, digitized speech, and recorded speech: A comparison of listener comprehension rates. *Proceedings of the Voice Input/Output Society*; Alexandria, VA. 1986.
- Murdock BB Jr. The serial position effect of free recall. *Journal of Experimental Psychology*. 1962; 64:482–488.
- Myers, JL.; Duffy, SA. Causal inferences and text memory. In: Graesser, AC.; Bower, GH., editors. *The Psychology of Learning and Motivation*, Vol. 25: Inferences and Text Comprehension. San Diego, CA: Academic Press; 1990. p. 159-173.
- Myers JL, Shinjo M, Duffy SA. Degree of causal relatedness and memory. *Journal of Memory and Language*. 1987; 26:453–465.
- Nusbaum, HC.; Dedina, JJ.; Pisoni, DB. Speech Research Laboratory Technical Note 84—02. Bloomington, IN: Indiana University; 1984. Perceptual confusions of consonants in natural and synthetic CV syllables.
- Nye, PW.; Gaitenby, JH. Haskins Laboratories Status Report on Speech Research, SR—33. New Haven, CT: Haskins Laboratories; 1973. Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results); p. 77-91.
- Nye, PW.; Gaitenby, JH. Haskins Laboratories Status Report on Speech Research, SR—37/38. New Haven, CT: Haskins Laboratories; 1974. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences; p. 169-190.
- Nye, PW.; Ingemann, F.; Donald, L. Haskins Laboratories Status Report on Speech Research, SR—41. New Haven, CT: Haskins Laboratories; 1975. Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms; p. 117-126.
- O'Brien EJ, Duffy SA, Myers JL. Anaphoric inference during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1986; 12:346–352.
- Pasdeloup, V. A prosodic model for French text-to-speech synthesis: A psycholinguistic approach. In: Bailly, G.; Benoit, C.; Sawallis, TR., editors. *Talking Machines: Theories, Models, and Designs*. Amsterdam: North-Holland; 1992. p. 335-348.
- Perfetti, CA. *Reading Ability*. New York: Oxford University Press; 1985.
- Perfetti, CA.; Roth, S. Some of the interactive processes in reading and their role in reading skill. In: Lesgold, AM.; Perfetti, CA., editors. *Interactive Processes in Reading*. Hillsdale, NJ: Erlbaum; 1981. p. 269-297.
- Pierce, L.; Remington, R. Comprehension of synthetic and natural speech in the presence of competing human speech. *Proceedings of the Voice I/O Systems Applications Conference '84*; September 11–13, 1984; Arlington, VA. 1984.
- Pisoni DB. Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America*. 1981; 70:S98.
- Pisoni, DB. Some measures of intelligibility and comprehension. In: Allen, J.; Hunnicutt, MS.; Klatt, DH., editors. *From Text to Speech: The MITalk System*. Cambridge, UK: Cambridge University Press; 1987. p. 151-171.
- Pisoni, DB.; Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. 1980 IEEE International Conference on Acoustics. Speech and Signal Processing; New York: IEEE; 1980. p. 572-575.
- Pisoni DB, Manous LM, Dedina MJ. Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*. 1987; 2:303–320.
- Pollack I. Message uncertainty and message reception. *Journal of the Acoustical Society of America*. 1959; 31:1500–1508.
- Pollack I. Message probability and message reception. *Journal of the Acoustical Society of America*. 1964; 36:937–945.

- Ralston, JV.; Lively, SE.; Pisoni, DB.; Rivera, S. Research on Speech Perception Progress Report No. 16. Bloomington, IN: Indiana University; 1990. Measuring the workload of comprehending spoken discourse: A first report; p. 155-167.
- Ralston JV, Pisoni DB, Lively SE, Greene BG, Mullennix JW. Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times. *Human Factors*. 1991; 33:471-491. [PubMed: 1835449]
- Rosson MB. Using synthetic speech for remote access to information. *Behavior Research Methods, Instruments, and Computers*. 1985; 17:250-252.
- Rundus D. Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*. 1971; 89:63-77.
- Salasoo A, Pisoni DB. Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*. 1985; 24:210-231.
- Samuel A. Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*. 1981; 110:474-494. [PubMed: 6459403]
- Schneider W, Shiffrin RM. Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*. 1977; 84:1-66.
- Schwab EC, Nusbaum HC, Pisoni DB. Some effects of training on the perception of synthetic speech. *Human Factors*. 1985; 27:395-408. [PubMed: 2936671]
- Shields JL, McHugh A, Martin JG. Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*. 1974; 102:250-255.
- Simpson, CA. Occupational Experience with a Specific Phraseology: Group Differences in Intelligibility for Synthesized and Human Speech. Paper presented at the ninetieth meeting of the Acoustical Society of America; San Francisco. 1975.
- Simpson, CA.; Navarro, T. Intelligibility of Computer Generated Speech as a Function of Multiple Factors. Paper presented at the National Aerospace and Electronics Conference; Dayton, OH. 1984.
- Slowiaczek LM, Nusbaum HC. Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*. 1985; 27:701-712. [PubMed: 2939018]
- Stanovich KE. Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*. 1980; 16:32-71.
- Stanovich KE, West RF. The effect of sentence context on ongoing word recognition: Tests of a two-process theory. *Journal of Experimental Psychology: Human Perception and Performance*. 1981; 7:658-672.
- Terken J, Lemeer G. Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics*. 1988; 16:453-457.
- Terken J, Nootboom SG. Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information. *Language and Cognitive Processes*. 1987; 2:145-163.
- Thibadeau R, Just MA, Carpenter PA. A model of the time course and content of reading. *Cognitive Science*. 1982; 6:157-203.
- Tyler LK, Warren P. Local and global structure in spoken language comprehension. *Journal of Memory and Language*. 1987; 26:638-657.
- West RF, Stanovich KE. Automatic contextual facilitation in readers of three ages. *Child Development*. 1978; 49:717-727.

Fig. 1.
Error rates (in percent) for various synthesis systems tested in both the closed- and open-response format MRT (from Logan, Greene, and Pisoni, 1989).

Fig. 2.
Response times (in msec) and percent correct for natural and synthetic words and non-words in a lexical decision task (from Pisoni, 1981).

Fig. 3.
Mean number of natural and synthetic words recalled as a function of memory preload (from Luce, Feustel, and Pisoni, 1983).

Fig. 4.
Number of subjects correctly recalling all of the digits as a function of memory preload
(from Luce *et al.*, 1983).

Fig. 5.
Probability of recall at each serial position for natural and synthetic word lists (from Luce *et al.*, 1983).

Fig. 6.
Percent correct for different categories of information presented in natural and synthetic speech (from Luce, 1981).

Fig. 7. Probability of a correct response for two kinds of information presented in natural and synthetic speech (from Ralston, Pisoni, Lively, Greene, and Mullennix, 1991).

Fig. 8. Sentence verification response times (in msec) for True and False responses to three- and six-word sentences presented in seven voices (from Manous, Pisoni, Dedina, and Nusbaum, 1985). Means are based on only those trials on which the subject verified and transcribed the sentence correctly.

Fig. 9. Mean sentence verification times (in msec) for True and False responses to three- and six-word sentences presented in natural and synthetic speech (from Pisoni, Manous, and Dedina, 1987). High-predictability sentences are displayed with open bars; low-predictability sentences are displayed with striped bars. Means are based on only those trials on which the subject verified and transcribed the sentence correctly.

Fig. 10. Sentence-by-sentence listening times as a function of voice and text (from Ralston *et al.*, 1991). Open bars represent natural speech; striped bars represent synthetic speech. Error bars represent one standard error of the sample mean.

Table 1

Mean signal duration (in msec) required for the identification of synthetic and natural target words in neutral and congruent contexts (from Duffy and Pisoni, 1991)

	Voice	
	Natural	Synthetic
Context		
Congruent	129 (3.3)	191 (5.0)
Neutral	335 (2.5)	408(12.5)

Note: Trials on which a subject never successfully identified the target word are excluded from the duration means. Percentages of trials on which subjects failed to identify the target words are presented in parentheses.