

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 26 (2003-2004)
Indiana University

Perception and Comprehension of Synthetic Speech¹

Stephen J. Winters and David B. Pisoni

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This research was supported by the NIH-NIDCD Research Grant R01 DC-00111 and NIH-NIDCD T32 Training Grant DC-00012 to Indiana University. We would like to thank Horabail Venkatagiri and Jan van Santen for their helpful comments and suggestions during the preparation of this paper.

Perception and Comprehension of Synthetic Speech

Abstract. An extensive body of research on the perception of synthetic speech carried out over the past 30 years has established that listeners have much more difficulty perceiving synthetic speech than natural speech. Differences in perceptual processing have been found in a variety of behavioral tasks, including assessments of segmental intelligibility, word recall, lexical decision, sentence transcription, and comprehension of spoken passages of connected text. Alternative groups of listeners—such as non-native speakers of English, children and older adults—have even more difficulty perceiving synthetic speech than young, healthy, college-aged listeners typically tested in perception studies. It has also been shown, however, that the ability to perceive synthetic speech improves rapidly with training and experience. Incorporating appropriate prosodic contours into synthetic speech algorithms—along with providing listeners with higher-level contextual information—can also aid the perception of synthetic speech. Listener difficulty in processing synthetic speech has been attributed to the impoverished acoustic-phonetic segmental cues—and inherent lack of natural variability and acoustic-phonetic redundancy—in synthetic speech produced by rule. The perceptual difficulties that listeners have in perceiving speech which lacks acoustic-phonetic variability has been cited as evidence for the importance of variability to the perception of natural speech. Future research on the perception of synthetic speech will need to investigate the sources of acoustic-phonetic variability and redundancy that improve the perception of synthetic speech, as well as determine the efficacy of synthetically produced audio-visual speech, and the extent to which the impoverished acoustic-phonetic structure of synthetic speech impacts higher-level comprehension processes. New behavioral methods of assessing the perception of speech by human listeners will need to be developed in order for our understanding of synthetic speech perception to keep pace with the rapid progress of speech synthesis technology.

Introduction

Studying the perception of synthetic speech has proven to be useful in many different domains. When work on the perception of synthetic speech first began in the early 1970s, researchers were primarily interested in evaluating its segmental intelligibility in comparison to natural speech. These early studies were done with an eye toward improving the quality of synthetic speech for use in practical applications, such as reading machines for the blind or voice output communication devices. Early evaluation studies such as Nye and Gaitenby (1973) assessed listener's perception of synthetic and natural speech not only to quantify the intelligibility gap between the two types of speech but also to isolate and attempt to identify which synthetic speech segments were difficult for listeners to perceive correctly. Once such segments were identified, further research could be carried out to improve the intelligibility of these segments through the refinement of the text-to-speech algorithms which produced them. In this way, researchers hoped to be able to improve the overall segmental intelligibility of synthetic speech to the same level as natural speech.

The initial studies by Nye and Gaitenby (1973) have led to a long line of research which has developed continually more sophisticated metrics for evaluating the intelligibility of synthetic speech. Standards for assessing the quality of text-to-speech synthesis have been proposed by Pols (1989; 1992) and van Santen (1994). Pols (1989) grouped together the various assessment techniques into four broad categories: 1. *global* techniques, addressing acceptability, preference, naturalness and usefulness; 2. *diagnostic* techniques, addressing segmentals, intelligibility and prosody; 3. *objective* techniques, including metrics such as the Speech Transmission Index (STI) and the Articulation Index (AI); and 4.

application-specific techniques, addressing the use of synthetic speech in specific applied domains such as reading machines and televised weather briefings. The bulk of early research on the perception of synthetic speech focused primarily on the assessment of segmental intelligibility (Type 2 techniques) and some global measures of preference and acceptability (Type 1 techniques).

Another group of researchers realized that much could be learned about speech perception by simply trying to figure out why people perceived synthetic speech differently than natural speech. Instead of asking how the intelligibility of synthetic speech might be improved, they asked “*why* is it that synthetic speech is more difficult to understand than natural speech? Does synthetic speech lack certain fundamental characteristics of natural speech which might be helpful to perception?” And, can listeners adjust or attune their perceptual systems to overcome such shortcomings in the synthetic speech signal?

The answers to these kinds of questions potentially lie much deeper than the surface level of segmental intelligibility. In order to study these problems, therefore, Pisoni (1981) suggested that more research on synthetic speech perception ought to be done in the ten areas shown in Table 1:

1. Processing time experiments
2. Listening to synthetic speech in noise
3. Perception under differing attentional demands
4. Effects of short- and long-term practice
5. Comprehension of fluent synthetic speech
6. Interaction of segmental and prosodic cues
7. Comparisons of different rule systems and synthesizers
8. Effects of naturalness on intelligibility
9. Generalization to novel utterances
10. Effects of message set size

Table 1. Needed research on the perception of synthetic speech (adapted from Pisoni 1981)

All of these lines of research have been pursued—to varying extents—in the years following Pisoni (1981). The results of this work have shown consistent differences in perception between natural and synthetic speech at every level of analysis (Duffy & Pisoni, 1992; Pisoni, 1997; Pisoni, Nusbaum & Greene, 1985).

Researchers have attempted to account for the perceptual differences between natural and synthetic speech in terms of the acoustic-phonetic characteristics which differ between these two types of speech. Throughout most of the ‘80s and ‘90s, research on synthetic speech perception used text-to-speech (TTS) systems which produced synthetic speech by rule. Synthesis-by-rule operates just as its name implies—a synthesis algorithm takes an orthographic string of letters as input and automatically converts them into speech output by using a set of text-to-speech conversion rules. Using rules to produce speech in this way results in a synthetic speech signal which lacks much of the variability inherent in natural speech. Furthermore, synthetic speech produced by rule typically provides fewer redundant cues to particular segments or sound contrasts. Text-to-speech conversion rules also tend to use highly simplified coarticulation sequences between individual segments, and they often lack appropriate or natural-sounding sentential prosody contours. Researchers have typically focused on these acoustic aspects of synthetic speech produced by rule in attempting to account for why listeners have greater difficulty perceiving it. Accounting for the difficulties of synthetic speech perception in this way has led researchers to draw some important conclusions about the nature of normal human speech perception—it apparently relies on redundant cues to particular segments, and it also makes use of the natural acoustic-

phonetic variability inherent in the speech signal. Redundancy and variability in the signal are fundamental properties of speech perception, and not just noise which needs to be filtered away through some kind of perceptual normalization process.

This chapter reviews evidence from the study of synthetic speech perception which has led researchers to draw these conclusions about normal speech perception. It details the various distinctions which have been found to exist between synthetic and normal speech perception, from the first segmental intelligibility studies to more global studies on comprehension, perceptual learning effects, alternative groups of listeners, and properties like naturalness and prosody. The final section concludes with some suggestions for potentially fruitful areas of future research in the context of the rapidly changing world of speech synthesis technology.

Point 1: Synthetic Speech is Less Intelligible than Natural Speech

The Modified Rhyme Test. Investigations of the segmental intelligibility of synthetic speech have routinely shown that it is less intelligible than natural speech. For instance, one of the first studies on the perception of synthetic speech, Nye and Gaitenby (1973), assessed the intelligibility of natural speech and synthetic speech produced by the Haskins Parallel Formant Resonance Synthesizer (Mattingly, 1968). They measured segmental intelligibility with the Modified Rhyme Test (MRT). The MRT, which was originally developed by Fairbanks (1958) and House, Williams, Hecker and Kryter (1965), presents listeners with a spoken word and then requires them to select the word they heard from a set of six, real-word alternatives, all of which differ by only one phoneme. Examples of these response sets include the following:

(1)		(2)	
game	came	dull	dub
fame	name	duck	dun
tame	same	dug	dud

Nye and Gaitenby (1973) adapted this task to the study of synthetic speech perception under the assumption that it would help them identify the particular segments for which the Haskins Pattern Playback produced poor, insufficient, or confusing cues.

In general, the results of Nye and Gaitenby's MRT study showed that synthetic speech was significantly less intelligible than natural speech. Listeners' overall error rate was 7.6% on the MRT for synthetic speech, but only 2.7% for natural speech. Nye and Gaitenby's results also revealed that synthetic obstruents were particularly unintelligible, as they induced error rates that were much higher than those for synthetic sonorants. However, Nye and Gaitenby pointed out that the MRT was of limited utility as an evaluative diagnostic for synthetic speech, since the closed response set provided a limited number of possible phonemic confusions and, moreover, it did not present all phonemes in equal proportions in the test stimuli.

Despite the shortcomings noted by Nye and Gaitenby (1973), the MRT has been used extensively in the ensuing years as virtually the standard way to assess the segmental intelligibility of various types of synthetic speech (Greene, Manous & Pisoni, 1984; Hustad, Kent & Beukelman, 1998; Koul & Allen, 1993; Logan, Greene & Pisoni, 1989; Mitchell & Atkins, 1988; Pisoni, 1987; Pisoni & Hunnicutt, 1980; Pisoni et al., 1985). Another problem with the original version of the MRT that Nye and Gaitenby used was that it was simply too easy. Ceiling effects are obtained in the MRT when natural or high-quality synthetic stimuli are presented under clear listening conditions. Hence, several studies have expanded

upon the MRT paradigm by preserving the original MRT word list but eliminating the forced-choice response set. In this “open-response format” version of the MRT, listeners simply respond to each particular word by writing down whatever they think they heard.

Logan et al. (1989) showed that this version of the MRT can help reduce the ceiling effects in performance that often emerge from the relative ease of the original forced-choice task. These authors used both the “closed-set” and “open-set” response formats of the MRT in testing the segmental intelligibility of 10 different speech synthesizers along with natural speech. For each synthesizer and each response format, Logan et al. tested 72 different listeners. Figure 1 shows the percentages of errors that these groups of listeners made in attempting to identify words produced by all of these voices, in both the open and closed formats of the MRT. Listener error rates were higher for all versions of synthetic speech than for natural speech; they were also higher in the open-set form of the task than they were in the closed-set format. These results indicate, moreover, that the intelligibility of synthetic speech depends greatly on the type of synthesizer being used to produce the speech. The low error rate observed for DECTalk in the closed response format ($\approx 5\%$), for instance, approximated that of natural speech, while the error rates for the worst synthesizers were consistently higher than 50% in the open response format.

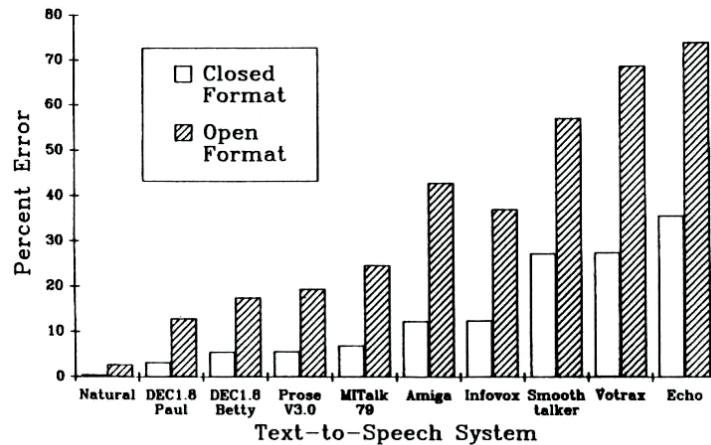


Figure 1. Error rates for word recognition in open and closed format MRT, by synthetic voice type (adapted from Logan et al. 1989).

Intelligibility of Synthetic Speech in Noise. The poor performance of most speech synthesizers in the open-response format in Logan et al. (1989) shows that the intelligibility gap between natural and synthetic speech increases significantly under more realistic testing situations. Offering listeners six possible response alternatives for each word they hear does not realistically reflect actual usage of speech synthesis applications, where listeners have to interpret each word they hear in terms of their entire lexicons. Likewise, testing the intelligibility of synthetic speech under ideal listening conditions in the laboratory does not reflect real-world usage either, where synthetic speech is often heard and produced in noisy environments. Several researchers have therefore tested the intelligibility of synthetic speech as it is played through noise and have found that such less-than-ideal listening conditions degrade the intelligibility of synthetic speech even more than they do the intelligibility of natural speech.

Pisoni and Koen (1981) were the first to report the results of testing synthetic speech in noise. They used the open- and closed-response formats of the MRT to test the intelligibility of natural and

synthetic (MITalk) speech in various levels of white noise. Figure 2 shows the percentages of words correctly identified in the different listening conditions in Pisoni and Koen's study. These results replicate earlier findings that intelligibility is worse for synthetic than for natural speech—especially in the open-response format of the MRT. Pisoni and Koen's results also established that increasing the level of noise had a significantly greater detrimental effect on the intelligibility of synthetic speech than it had on the intelligibility of natural speech.

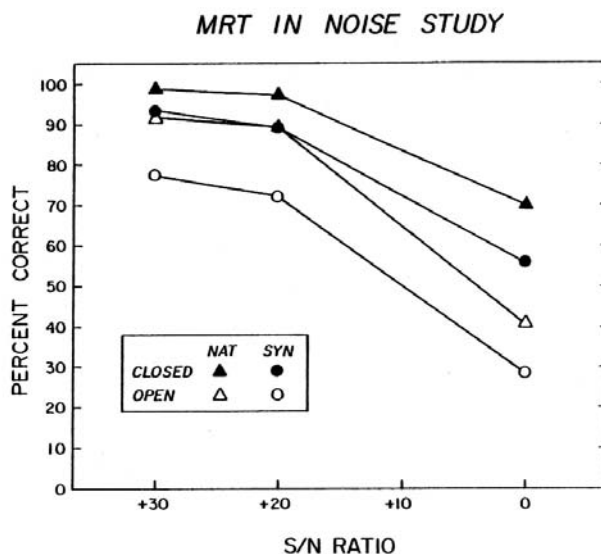


Figure 2. Percent correct word recognition in open and closed-format MRT, by voice type and signal-to-noise ratio (adapted from Pisoni & Koen, 1981).

In another study, Clark (1983) also tested the perception of synthetic speech, as produced by his own synthesis algorithms, in white noise. Instead of using the MRT, however, Clark measured the perception of vowels in /hVd/ sequences and consonants in /Ca:/ syllables, in an attempt to assess the susceptibility of particular segmental cues to degradation in noise. He found that the intelligibility of his synthetic vowels was surprisingly robust in noise, even to the point of being marginally more intelligible than natural vowels in the noisiest listening conditions (-6 db SNR). Clark did, however, find appreciable degradation of the intelligibility of synthetic consonants in noise, especially stops and fricatives. Despite the comparatively greater degradation of synthetic consonant intelligibility in noise, however, Clark found that the rank-ordering of individual consonant intelligibilities did not differ significantly between natural and synthetic speech in the noisy listening conditions. Clark therefore concluded that synthetic speech might be considered a “degraded” form of natural speech, with cues to individual segments that were simply not as salient as those found in natural speech.

Nusbaum, Dedina and Pisoni (1984) took issue with Clark's assessment and argued that the segmental cues in synthetic speech were not just “degraded” forms of the ones found in natural speech, but that they were “impoverished” and could also be genuinely misleading to the listener. In support of this claim, Nusbaum et al. presented results from a perception experiment in which they played CV syllables, produced by both a human voice and three different speech synthesizers, to listeners in various levels of noise. When Nusbaum et al. investigated the confusions that listeners made in trying to identify the consonants in this experiment, they found that listeners not only had more difficulty identifying synthetic consonants in noise, but that they also often misidentified synthetic consonants in ways that

were never seen in the confusions of natural speech tokens. Even in listening to DECTalk speech, for instance—a form of synthetic speech that has been established as highly intelligible by studies like Logan et al. (1989)—listeners often misidentified /r/ as /b/—a perceptual confusion that never occurred for natural speech tokens of /r/. Nusbaum et al. therefore suggested that some synthetic speech cues could be misleading to the listener as well as being impoverished versions of their natural counterparts.

More recent investigations of the perception of synthetic speech in noise have incorporated multi-talker “babble” noise into the signal, rather than white noise. This methodology captures another level of ecological validity, since natural conversations—and the use of synthetic speech applications—often occur in public, multi-talker settings, where it is necessary to focus in on one particular voice through the “cocktail party effect” of other speakers (Bregman, 1990). Studies which have incorporated this type of noise in tests of synthetic speech intelligibility have found that it does not generally degrade the intelligibility of synthetic speech as much as white noise. Koul and Allen (1993), for instance, incorporated multi-talker noise into an open-response version of the MRT, using both natural and DECTalk voices. While they found that the natural voice was more intelligible than DECTalk—and that both voices were more intelligible at the higher signal-to-noise ratios (SNRs)—they did not find any interaction between voice type and the level of multi-talker babble noise. Both voices, that is, suffered similar losses in intelligibility at decreasing SNRs. For instance, at a +25 db SNR, listeners correctly identified 84% of natural words and 66% of synthetic words, while at a 0 db SNR, correct word identification scores decreased to 52% for natural words and 30% for the DECTalk items. Koul and Allen also found similar identification errors for both natural and synthetic segments. Different confusions emerged for /h/, however, which was often misidentified in DECTalk, but rarely confused in natural speech. Koul and Allen suggested that the “babble” noise might influence the perception of synthetic speech in this unique way because its acoustic energy is primarily focused in the lower frequencies, rather than evenly spread across the spectrum as in white noise. Aside from this particular difference, Koul and Allen’s results indicated a high degree of similarity between the segmental cues used in DECTalk and those found in natural speech.

Perception of formant vs. concatenative synthesis. In recent years, speech synthesis technology has relied increasingly on “concatenative” synthesis as opposed to the “formant-based” synthesis techniques that were prevalent in the ‘70s and ‘80s (Atal & Hanauer, 1971; Dutoit & Leich, 1993; Moulines & Charpentier, 1990). Formant-based synthesis operated on the basis of an electronically implemented, source-filter model of the human articulatory system (Fant, 1960). These synthesizers produced speech by progressing through a series of source-filter targets, one segment at a time. This approach therefore focused on the acoustic quality of the synthetic speech within each individual segment, rather than on the acoustic transitions between the segments. Concatenative synthesis, on the other hand, uses an actual human voice as its source—rather than the output of an electronic model—and incorporates segmental boundaries within its basic unit of production. The size of the basic units used in concatenative synthesis may be as large as a sequence of words or as small as a portion of a phoneme, but a popular choice is a diphone-sized unit, which extends from the midpoint of one phoneme to the midpoint of the next phoneme. Each such diphone thereby contains the acoustic transition between a particular pair of phonemes. Concatenative synthesis algorithms operate by simply joining these basic diphone units together into the desired phonemic string.

Encoding the transitions between segments into the speech output—along with using a natural human voice as the speech source—is supposed to make concatenative synthesis sound more natural and aesthetically appealing to the listener than formant-based synthesis. While these aspects of concatenative synthesis may have motivated its increasing popularity, there has actually been little research demonstrating that it is either more intelligible or even more natural-sounding than formant-based synthesis. Studies testing early forms of diphone-based synthesis, such as RealVoice and SmoothTalker,

indicated that they were consistently less intelligible than high-quality formant synthesis systems, such as DECTalk (Logan et al. 1989). Subsequent studies have shown, however, that more recent diphone-based synthesizers can match DECTalk in intelligibility. Rupprecht, Beukelman and Vrtiska (1995) tested the comparative intelligibility of DECTalk versus MacinTalk (an early form of diphone-based synthesis) and MacinTalk Pro (an improved version of MacinTalk). Rupprecht et al. played listeners sentences from the Speech Perception in Noise (SPIN) test (Kalikow, Stevens & Elliott, 1977) as produced by each of the synthetic voices, and asked listeners to identify the last word in each sentence. Rupprecht et al. found that correct identification rates for both DECTalk and MacinTalk Pro were significantly higher than the same rates for the MacinTalk synthesizer. There were, however, no significant differences between the intelligibility of the DECTalk and MacinTalk Pro voices. Hustad et al. (1998) also tested the intelligibility of DECTalk and MacinTalk Pro synthesis in an open-response format MRT. While the intelligibility of both voices was quite high in this task, Hustad et al. found a slight but significant advantage in intelligibility for the DECTalk voice. Hustad et al. suggested that Rupprecht et al.'s failure to find such a difference in their earlier study may have been the result of their listeners relying on higher-level contextual cues to identify words in the SPIN sentences.

In a more recent study, Venkatagiri (2003) also used words from the MRT to investigate the comparative intelligibility of four different synthesizers, each using a different combination of formant and concatenative synthesis in their speech production algorithms. These included AT&T's NextGen TTS, which uses half-phone based synthesis (a method of concatenating halves of individual phonemes together); Festival, which uses diphone-based synthesis; FlexVoice, which uses a combination of formant and diphone-based synthesis; and IBM ViaVoice, which uses formant-based synthesis only. Venkatagiri (2003) tested the intelligibility of each of these systems—along with a human voice—by playing productions of individual words from the MRT in neutral-sentence carrier phrases (e.g., “The word is ____.”) in two different levels of multi-talker babble noise. Venkatagiri (2003) found that the listeners' ability to identify the natural speech tokens under these conditions was significantly greater than their ability to identify the same tokens as they were produced by all four of the synthetic voices. He also found that increasing the level of noise was significantly more detrimental to the intelligibility of all types of synthetic speech than it was for natural speech.

Among the four synthetic voices, the two voices that used concatenative synthesis techniques (NextGen and Festival) were significantly more intelligible than the one using formant-based synthesis (ViaVoice). In addition, FlexVoice, which used a combination of formant and concatenative algorithms was significantly less intelligible than the other three synthetic voices. Interestingly, formant-based synthesis outperformed the concatenative synthesizers on vowel intelligibility, even though it induced significantly more listener errors in the identification of consonant sounds. These results suggested that the ability of concatenative synthesizers to produce highly intelligible consonant sounds—presumably because they maintain the natural transitions between these transient segments and their surrounding phonemes—comes at the cost of being able to produce highly intelligible vowel sounds. Moreover, the relatively poor performance of the FlexVoice synthesizer—which combined the formant and concatenative approaches—indicated that low cost speech synthesis technology has not yet advanced to the point where it is possible to combine the best perceptual aspects of both production algorithms into one system. Venkatagiri (2003) also observed that the poor intelligibility of all synthetic voices in noisy conditions—no matter what their method of production—revealed that there are still serious limitations on their potential utility in real-world applications, where noisy listening conditions are the norm.

Summary. The available evidence suggests that the segmental intelligibility of synthetic speech is significantly worse than that of natural speech. Synthetically produced words have routinely been shown to be more difficult for listeners to identify in forced-choice tests of segmental intelligibility, such as the MRT. The gap between natural and synthetic speech intelligibility also increases substantially

when noise is introduced into the speech signal, or if there are fewer constraints on the possible number of responses listeners can make in a particular testing paradigm. The poor segmental intelligibility of synthetic speech produced by rule appears to result from the use of segmental cues which are not only acoustically degraded relative to natural speech but also impoverished and therefore potentially misleading to the listener. The use of naturally produced speech segments in concatenative speech synthesis techniques may provide a method of overcoming such intelligibility limitations, but research on the perception of synthetic speech produced in this manner indicates that it is still not as intelligible as natural speech, especially in adverse listening conditions.

Point 2: Perception of Synthetic Speech Requires More Cognitive Resources

Lexical Decision. One consequence of the poor segmental intelligibility of synthetic speech is that listeners may only be able to interpret it by applying more cognitive resources to the task of speech perception. Several studies have shown that listeners do, in fact, engage such compensatory mechanisms when they listen to synthetic speech. Pisoni (1981), for instance, had listeners perform a speeded lexical decision task, in which they heard both naturally produced and synthetically produced (MITalk) strings as test items. These test strings took the form of both words (e.g., “colored”) and non-words (e.g., “coobered”). Pisoni found that listeners consistently took more time to determine whether or not the synthetic strings were words than if the natural strings were words, regardless of whether or not the test string was a real lexical item or a non-word. Since no significant interaction emerged between the voice type and the lexical status of the string, Pisoni concluded that listeners had to apply more cognitive resources to the task of interpreting the acoustic-phonetic surface structure of synthetic strings, prior to any higher-level lexical or semantic processing.

In a follow-up study, Slowiaczek and Pisoni (1982) suggested that the processing advantage for natural speech in the lexical decision task might be the result of greater listener familiarity with natural speech. They assessed whether the processing gap between natural and synthetic speech items in a speeded lexical decision task might be reduced as listeners became more familiar with the synthetic voice. They investigated this possibility by having listeners perform a speeded lexical decision task, as in Pisoni (1981), in which listeners heard both word and non-word strings as produced by both a natural voice and MITalk. Slowiaczek and Pisoni’s listeners performed a speeded lexical decision task for five consecutive days, while listening to word and non-word strings produced by both a natural voice and MITalk. Figure 3 compares the response times for Slowiaczek and Pisoni’s listeners, on the fifth day of performing this task, to the response times from Pisoni’s (1981) listeners, on the only day on which they performed an identical task. Slowiaczek and Pisoni found that listener response times (RTs) decreased over the course of the five-day training process only for the strings produced by the synthetic voice; the RTs for the naturally produced items remained constant over time. The results in Figure 3 also show that, despite this improvement in performance, RTs for synthetic words never reached the same level as RTs for natural words (although RTs for both synthetic and natural non-words were quite close, after five days of testing). Slowiaczek and Pisoni concluded that the advantage for natural speech found in Pisoni (1981) reflected genuine differences in the processing of natural and synthetic speech, which could not be eliminated completely just by increasing listeners’ familiarity with synthetic speech.

AUDITORY LEXICAL DECISION

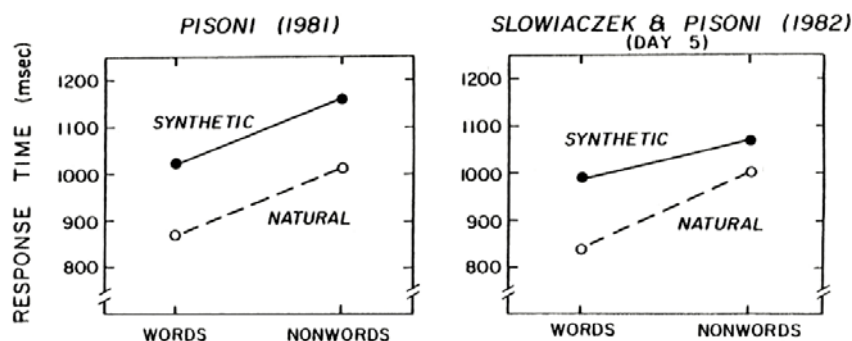


Figure 3. Lexical decision reaction times, for synthetic and natural words and non-words (adapted from Slowiaczek & Pisoni, 1982).

Word Recognition. In another study on spoken word recognition, Manous and Pisoni (1984) also found that synthetic speech puts greater demands on listeners in a word recognition task than natural speech. Using the gating paradigm developed by Grosjean (1980), Manous and Pisoni presented listeners with increasing amounts of individual words in a sentential context, as produced by both a human speaker and the DECTalk synthesizer. The amount of the word that each listener heard increased by 50 milliseconds on each successive trial; i.e., the listeners first heard 50 milliseconds of the word, then 100 milliseconds, and so on. Manous and Pisoni found that listeners needed to hear, on average, 361 milliseconds of naturally produced words before they could reliably identify them, whereas they needed to hear an average of 417 milliseconds of DECTalk-produced words before they could reach the same level of accuracy. Manous and Pisoni attributed this difference in performance to the acoustically impoverished nature of synthetic speech. Since DECTalk provides fewer redundant acoustic-phonetic cues for individual segments than natural speech does, listeners needed to hear and process more segmental information before they could reliably identify spoken lists of synthetic words.

Word Recall. The recall of synthetically produced words also requires more cognitive resources than the recall of naturally produced words. Luce, Feustel and Pisoni (1983) tested listeners' ability to recall unordered lists of words that were produced by both a human voice and MITalk. The authors found that listeners could recall more items from the natural word lists than they could from the synthetic word lists. They also found that there were more intrusions (words "recalled" by listeners that were not in the original lists) in the recall of the synthetic lists than in the recall of the natural lists. In a follow-up experiment, Luce et al. also measured listeners' ability to recall synthetic and natural word lists when they had a digit pre-load. Listeners first memorized a list of 0, 3 or 6 digits before they heard a list of 15 test words. The listeners were then asked to recall both lists—first the digits, in order, and then the words, in no particular order. Once again, the listeners recalled fewer words from the synthetic lists correctly than they did from the natural lists, and they also produced more "intrusions" in the recall of synthetic lists. Furthermore, the listeners' recall of the six-digit lists was worse when they were presented before lists of synthetic words. Luce et al. claimed that this interaction indicated that both digit and word recall shared the same, limited store of cognitive processing resources—and that the storage and maintenance of synthetic words absorbed more of these processing resources than the storage and maintenance of natural words.

Lastly, Luce et al. (1983) tested the serial recall of 10-word lists, produced by both natural and synthetic voices. In a serial recall task, listeners must recall a list of words in the correct order; this task thus requires the encoding of both item and order information. Luce et al. found both recency and primacy effects when listeners attempted to recall lists of synthetic and natural words in order; that is, they were best at recalling words that appeared both early and late in the various lists. However, these effects interacted with the type of voice used. While the recall of natural items was, in general, better overall than the recall of synthetic words, this advantage was significantly larger for early items than late items in the lists. Luce et al. hypothesized that this interaction might occur because the storage of synthetic words required more processing resources than the storage of natural words. Late items in the synthetic lists might therefore make use of memory resources that would otherwise be reserved for the storage of early items in the list. Thus, the recall of early synthetic items was disadvantaged not only by the fact that they required more memory resources than the early natural items, but also by the fact that subsequent items in the synthetic lists required more memory resources, as well.

Luce and Pisoni (1983) tested these hypotheses in another recall study, using mixed lists of synthetic and natural words. Following Luce et al.'s (1983) earlier logic, Luce and Pisoni hypothesized that synthetic words that appeared late in a list should adversely affect the recall of both natural and synthetic words that appeared earlier in the same list. Luce and Pisoni had listeners recall mixed lists in which five natural items were followed by five synthetic items (and vice versa). The results of this study did not support the prediction that the recall of either early synthetic or natural words would be hampered by late synthetic items. Luce and Pisoni's results also failed to replicate Luce et al.'s (1983) finding that the recall of early synthetic items was worse than the recall of early natural items. In order to determine whether or not these findings were the result of the poor intelligibility of the items in the synthetic word lists, Luce and Pisoni constructed new lists of both natural and synthetic words using only items that listeners identified correctly more than 98% of the time in an MRT task. Using these lists, Luce et al. found that the recall of natural words was better than the recall of synthetic words in positions 2, 3 and 5 of the 10 word-long lists; there were no significant differences in recall between natural and synthetic voices for the other positions in the word lists. Since differences in recall were found using lists of highly intelligible words whose acoustic-phonetic interpretation presumably required minimal amounts of extra cognitive effort, Luce and Pisoni concluded that the recall and higher-level processing of synthetic words genuinely did require more cognitive resources than the recall of natural words.

Summary. Evidence from lexical decision, word recognition and word recall studies suggest that the perception of synthetic speech requires more cognitive resources than the perception of natural speech. Listeners take longer to decide if synthetic strings are words in a lexical decision task, and they also need to hear more of a synthetic word before they can recognize it in a gated word recognition task. The existence of these processing deficits indicates that listeners must apply extra cognitive resources to the interpretation of the impoverished acoustic-phonetic cues of synthetic speech at the segmental level. Research showing that the storage and recall of synthetically produced words is more difficult than the storage and recall of naturally produced words also indicates that additional cognitive resources must be used in the encoding of synthetic speech items in memory.

Point 3: Perception of Synthetic Speech Interacts with Higher-Level Linguistic Knowledge

Perception of Words in Sentences and in Isolation. Listeners also compensate for the poor segmental intelligibility of synthetic speech by relying on any available higher-level linguistic information to help them interpret a synthetic speech signal correctly. For this reason, the perception of synthetic words presented in sentential contexts has been found to be significantly better than the perception of synthetic words in isolation.

Hoover, Reichle, Van Tasell and Cole (1987) for example, demonstrated the influence of higher-level linguistic information on the perception of synthetic speech by comparing listeners' perception of synthetically produced words in isolation to the perception of synthetically produced "low probability" and "high probability" sentences. "Low" and "high" probability sentences were used as a means of testing the effects of semantic plausibility on the perception of words in a sentential context. Hoover et al. constructed sentences of these two types by asking participants, in a pre-test, to fill in the final word in a series of short, declarative sentences. The final words that the participants chose to fit these contexts more than 90% of the time was used in the high-probability sentences, whereas the final words that were chosen sparingly were used in the low-probability sentences. Hoover et al. recorded sentences of these two types, along with the final words for each sentence in isolation, as produced by a human speaker and both the Votrax and the Echo II synthesizers. They then presented these words and sentences to listeners, who were instructed to repeat what they had heard. Hoover et al. found that the listeners repeated sentences of both types more accurately than they repeated the individual words. Repetition accuracy for the synthetic items was still worse than the repetition accuracy for the natural words and sentences, however. Table 2 shows the correct identification rates for all voices and conditions in Hoover et al.'s study:

	Single Words	Low-Prob. Sentences	High-Prob. Sentences
Votrax	21.9	35.3	87.8
Echo II	19.5	23.8	77.3
Natural	99.9	100	100

Table 2. Percentage of words correctly repeated, by voice type and presentation context (adapted from Hoover et al., 1987).

Hoover et al. (1987) also found that Votrax speech was more intelligible than Echo II speech in the two different sentence contexts. In general, however, the intelligibility of both of these synthesizers was quite low. The correct identification rates for words in the high probability sentences produced by these synthesizers was less than 90%, even though the words in these contexts were chosen on the basis of their being selected more than 90% of the time by readers who were merely filling in the blanks at the ends of these sentences. Such results suggest that poor-quality synthetic speech may actually mislead a listener and thereby be less informative to a listener than no signal at all.

Mirenda and Beukelman (1987) undertook a similar investigation of the perception of synthetic words in isolation and in sentences, but they included tokens produced by the highly intelligible DECTalk synthesizer, along with the poorer quality Echo and Votrax synthesizers. These authors found that correct identification rates for DECTalk increased from 78% in isolated word contexts to 96.7% in sentences, for adult listeners. This approximated adult listener performance on natural speech versions of the same tokens, which reached ceiling levels of performance at 99.2% correct for individual words and 99.3% correct for words in sentences. Mirenda and Beukelman (1990) expanded the range of synthesizers used to produce the stimuli in a follow-up study, which used the same methodology, and found once again that the sentential contexts improved the intelligibility of all synthetic voices. The best synthesizer in this second study, SmoothTalker 3.0, was a diphone-based system, but it still failed to reach the intelligibility levels found with natural speech in both isolated words and sentential contexts.

Semantically Anomalous Sentences. Earlier research on the intelligibility of synthetic words in sentences suggests, however, that the perception of words in sentences actually becomes worse than the perception of words in isolation if the contents of the sentences lack semantic coherence and

predictability. Nye and Gaitenby (1974), for instance, tested the intelligibility of the Haskins Parallel Formant Resonance Synthesizer with both the closed-set MRT and a set of syntactically normal but meaningless sentences (e.g., “The safe meat caught the shade.”) Nye and Gaitenby found that correct identification rates were much lower for synthetic words in these meaningless sentences (78%) than they were for the words presented in isolation in the MRT (92%). This effect was also proportionally greater for synthetic speech than it was for natural speech, which scored a 97% correct identification rate in the MRT and a 95% correct identification rate in the meaningless sentence condition. Nye and Gaitenby attributed the detrimental effect of sentential contexts to the difficulty listeners had in parsing individual words out of a longer string of words. They also pointed out that higher-level sentential information might have biased listeners towards expecting to hear words which fit into the sentence’s semantic context, rather than the words that did appear in the anomalous sentences that were presented to them.

Pisoni and Hunnicutt (1980) presented further evidence to support the hypothesis that semantically anomalous sentences produced detrimental effects on the perception of words. They tested listeners on the closed-set MRT, using both natural speech and the MITalk system. The listeners were also asked to transcribe both a set of semantically meaningless Haskins sentences and a set of meaningful Harvard sentences (Egan, 1948). Pisoni and Hunnicutt found that correct identification rates for individual words were comparable between the MRT (99.4% for natural speech, 93.1% for synthetic) and the Harvard sentences (99.2% for natural speech, 93.2% for MITalk), but were lower for the semantically meaningless Haskins sentences (97.3% for natural speech, 78.7% for synthetic speech). The loss of meaning in sentential contexts thus had a more detrimental effect on the intelligibility of synthetic speech than it did on the intelligibility of the natural voice.

Gating in Sentences. Duffy and Pisoni (1991) noted that the correct identification rates for both natural and DECTalk speech in Mirenda and Beukelman’s (1987) tests of sentence transcription were close to ceiling levels of performance. They therefore developed a gating paradigm for words in sentential contexts in an attempt to tease apart the similar intelligibility levels of these two kinds of speech. Duffy and Pisoni’s gating paradigm involved presenting sentences to listeners in which they heard increasing amounts of the final word on successive trials. On the first presentation, the listeners heard none of the final word; on the second presentation, they heard 50 ms of the word; on the third, they heard 100 ms, and so on. After each of these presentations, the listeners were instructed to guess what the final word in the sentence was. Duffy and Pisoni presented these words to the listeners in either a “congruent” or a “neutral” sentential context, using both DECTalk and human voices. The final words in the congruent sentences were semantically related to the words at the beginning of the sentences (e.g., “The soldiers flew in the helicopter.”) while the final words in the neutral sentences had no clear semantic relation to the words at the beginning of those sentences (e.g., “The people were near the helicopter.”)

Figure 4 shows the percentage of correct identifications of a word, in both natural and synthetic voices, for each presentation of that word at the various gating durations. On average, listeners needed to hear 68 ms more of the words if they were produced by DECTalk than if they were spoken in a natural voice before they could reliably identify them in these contexts. This effect of voice type interacted significantly with the type of sentence context; listeners had to hear 212 ms more of the synthetic words when they appeared in the “neutral” contexts than when they appeared in the “congruent” contexts before they could identify them reliably. These results indicated, once again, that the perception of synthetic speech not only requires more cognitive resources than the perception of natural speech, but also that listeners appear to draw much more heavily on higher-level syntactic and semantic information in order to compensate for the difficulties they incur in processing the impoverished acoustic-phonetic structure of synthetic speech.

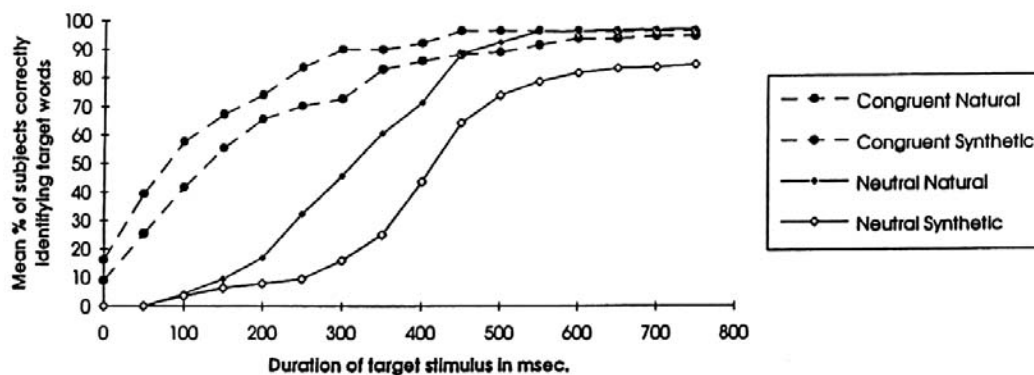


Figure 4. Percent of synthetic and natural words correctly identified, by duration of target stimulus in congruent and neutral sentential contexts (adapted from Duffy & Pisoni, 1991).

Summary. Research on the perception of synthetically produced sentences indicates that higher-level syntactic and semantic information can both facilitate and hinder the perception of synthetic speech. The transcription of synthetic words in meaningful sentences is typically better than the identification of those words in isolation. This finding suggests that listeners rely more extensively on the higher-level linguistic information in meaningful sentences to help them overcome difficulties they may have in interpreting the synthetic speech signal at the segmental level. However, listeners also have more difficulty identifying synthetic words in semantically anomalous sentences than they do identifying synthetic words in isolation, perhaps because the higher-level semantic information in these sentences encourages the listeners to develop misleading expectations about which words may follow. Semantic information has also been shown to interact with the perception of individual words at the ends of sentences in a gating paradigm, indicating that listeners may simultaneously apply both their knowledge of higher-level linguistic structure and additional cognitive resources to the challenging task of interpreting the impoverished acoustic-phonetic structure of synthetic speech.

Point 4: Synthetic Speech is More Difficult to Comprehend than Natural Speech

Despite the wealth of evidence from many studies showing that synthetic speech is less intelligible than natural speech, early investigations of listeners' ability to comprehend synthetic speech often showed that it was no worse than their ability to comprehend natural speech. Many of these early studies, however, relied on post-perceptual measures of comprehension, using multiple-choice questions or free recall tasks. These studies assessed the "product" rather than the "process" of language comprehension. The results of these studies may have, therefore, reflected the influence of other, higher-level cognitive processes which extend far beyond the scope of perceiving and processing the synthetic speech signal itself.

Post-perceptual Comprehension Tests. Nye, Ingemann and Donald (1975) were the first researchers to study listeners' comprehension of synthetic speech. They played college-level reading passages to listeners and then asked them to answer a series of multiple-choice questions based on those passages. The listeners were instructed to take as much time as they needed to get as many of the questions right as they possibly could. They could even go back and re-play sections of the passages that they had difficulty understanding the first time. The dependent measure that Nye et al. looked at in this paradigm was the total amount of time it took listeners to answer the multiple-choice questions. Nye et al. found that this amount of time was significantly longer when the passages were played to the listeners using synthetic speech generated by the Haskins Parallel Resonant Synthesizer than when they were

played using natural speech. The proportion of questions that listeners answered correctly did not vary according to the type of speech used, however.

Subsequent studies on the comprehension of continuous, connected passages of synthetic speech have yielded a similar pattern of results: it may take longer to process synthetic speech than natural speech, but the final levels of comprehension achieved for both types of speech are ultimately equivalent. In a replication of Nye et al.'s (1975) earlier study, Ingemann (1978) used the more advanced FOVE speech synthesizer and found no differences in performance between natural and synthetic voices in either the amount of time listeners took to complete the task or the number of multiple-choice questions they answered correctly. Pisoni and Hunnicutt (1980) assessed comprehension by having listeners answer multiple-choice questions based on passages that they had either read or heard spoken by a human voice or the MITalk text-to-speech system. They found that listeners answered more questions correctly if they had read the passages (rather than heard them), but that their level of performance did not differ between the natural voice and MITalk conditions. They also found that the percentage of questions that participants answered correctly improved most between the first and second halves of the experiment when the participants heard the MITalk versions of the passages, suggesting that some perceptual learning of this synthetic voice had occurred during the course of the experiment. Participants made similar, but smaller amounts of improvement between the two halves of the study in the natural voice and reading conditions.

Online Measures: Sentence Verification. Conclusive evidence showing that the comprehension of synthetic speech was more difficult for listeners than the comprehension of natural speech first began to emerge when researchers started applying more sensitive, online measures to the study of the comprehension process. Manous, Pisoni, Dedina and Nusbaum (1985) were the first researchers to use a sentence verification task (SVT) to study the comprehension of synthetic speech. They played listeners short, declarative sentences, which were either verified or falsified by the last word in the sentence and instructed the listeners to decide as quickly and as accurately as possible whether each sentence was true or false. The listeners recorded their responses by pressing one of two buttons on a response box and then writing down the sentence they heard. Manous et al. presented sentences to their listeners in this experiment using four different synthetic voices: DECTalk, Prose, Infovox and Votrax, as well as a human voice. They found that listener reaction time in the SVT was related to the intelligibility of the speech used, as determined by the number of transcription errors listeners made in writing down the sentences they heard. Manous et al. therefore concluded that the difficulty of encoding synthetic speech at the acoustic/phonetic level had a cascading effect on higher levels of the comprehension process.

In another study, Pisoni, Manous and Dedina (1987) followed up on Manous et al.'s (1985) earlier findings by investigating whether the increased latencies for synthetic speech in the SVT were due to the impoverished segmental cues in synthetic speech, or if they also resulted from independent difficulties the listeners incurred in processing synthetic speech beyond the level of acoustic-phonetic surface structure. Pisoni et al. assessed this hypothesis by testing listeners in a SVT using stimuli produced by human and synthetic voices that were matched in terms of segmental intelligibility. Pisoni et al. thus replicated the methodology from Manous et al. using a DECTalk voice that did not induce significantly more errors than a natural voice in a sentence transcription task. Figure 5 shows the average response latencies from this study, for true and false sentences, broken down by the type of voice used and the length of the sentence (three or six words). The reaction times for true sentences were significantly longer for the DECTalk voice than for the natural voice—despite the close match between the two voices in terms of segmental intelligibility. Pisoni et al. thus concluded that the comparatively impoverished acoustic-phonetic structure of DECTalk had detrimental effects on the comprehension process that did not emerge at the level of segmental encoding. In fact, this finding suggested that a level of processing devoted strictly to the segmental encoding of incoming speech might not even exist, since

the ramifications of processing impoverished acoustic-phonetic cues seemed to persist beyond the segmental encoding stage. Pisoni et al.'s findings also suggested that the higher-level structures produced by the comprehension process were more impoverished—or more difficult to interpret—for the synthetic speech stimuli than they were for the natural speech stimuli, since the listeners' longer reaction times in the SVT did not correspond to any difficulties observed at the earlier segmental processing level.

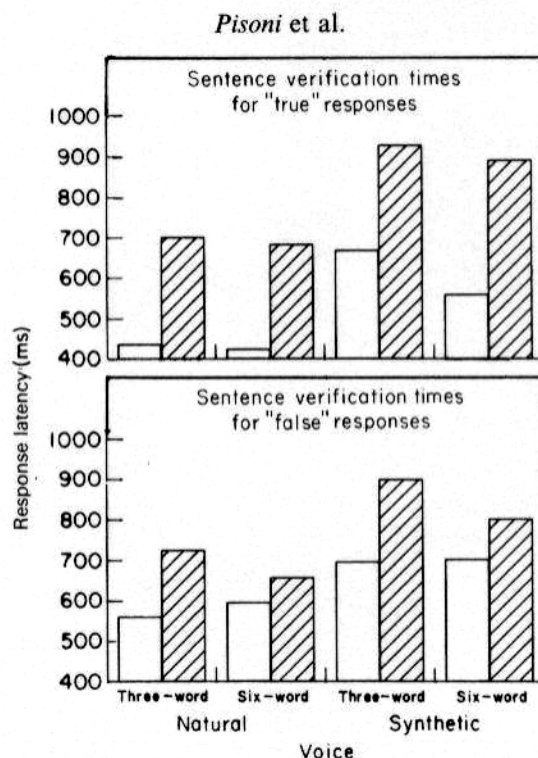


Figure 5. Response latencies for verification of false and true sentences, by voice type and length of sentence (adapted from Pisoni, Manous & Dedina, 1987).

Comprehension and Recall. Luce (1981), however, suggested that the difficulty of interpreting the acoustic-phonetic structure of synthetic speech may actually make the subsequent recall of synthetic speech easier than the recall of naturally produced speech. Luce drew upon this notion to account for a finding which showed that listener recall of individual words was better when they were produced by MITalk than when they were produced by a natural voice. Luce suggested that this followed from the extra cognitive effort listeners needed to make in order to encode synthetic speech successfully, which evidently resulted in a stronger memory trace for the particular acoustic details of the words the listener heard. Luce based this prediction on the results of a study in which listeners heard passages spoken in either MITalk or a natural voice. Luce then questioned the listeners about whether particular words, propositions or themes had appeared in those passages. Luce found that listeners were able to recall particular propositions and themes better when they heard the passage in a natural voice, rather than in synthetic speech. However, the listeners' memory for particular words was better when they had heard passages produced by the MITalk voice. Luce suggested that better recall for individual synthetic words followed from the fact that listeners had to expend greater amounts of cognitive effort and processing resources on encoding the acoustic-phonetic details of the synthetic speech into words. Allocating more

cognitive effort to the process of word recognition, as it were, led to more robust lexical recall but produced poorer recall of more abstract propositional information.

Moody and Joost (1986) also measured the recall of words and propositions in synthetic and natural passages and found that recall interacted in unexpected ways with higher-level conceptual structures. They asked listeners multiple-choice questions from college and graduate entrance exams about passages which had been presented to them in DECTalk, LPC synthesis, and natural voices. Moody and Joost found that listeners answered more multiple-choice questions correctly when those questions dealt with topics in the naturally produced passages, rather than the synthetic ones. However, this difference in comprehension only held for the easier, or less complicated, propositions in the original passage; no differences in percent correct response were found for questions about more difficult passages. Their findings suggested once again that expending greater amounts of cognitive effort in interpreting the more difficult portions of a synthetically produced passage may help listeners recall those propositions just as well as their naturally produced counterparts later on. Their results also suggest that, as the comprehension and recall tasks become more difficult, post-perceptual measures of comprehension may be influenced more by alternate processing strategies and knowledge from beyond the immediate scope of the task.

Ralston, Pisoni, Lively, Greene and Mullennix (1991) further investigated the relationship between synthetic speech intelligibility and comprehension by having listeners perform a series of tests which used natural speech and Votrax synthetic speech stimuli. Their listeners first took an MRT, then had to monitor for a small set of words in a short passage, and then finally answered a series of true/false questions regarding the presence or absence of words and propositions in the passage they had just heard. In the word monitoring task, Ralston et al. found that accuracy was higher and response times lower for natural speech than for Votrax speech. Interestingly, word-monitoring latencies increased significantly for the more difficult (i.e., college-level) passages only for synthetic speech, indicating that the process of recognizing individual synthetic words shared processing resources with higher-level comprehension processes. However, in contrast to Luce (1981), Ralston et al. found that listeners' memory for both words and propositions was better when they had been presented in natural speech than when they had been produced by the Votrax synthesizer. Hence, despite the extra cognitive effort required to encode the phonetic details of these synthetic stimuli, they were still more difficult to recall from memory.

Ralston et al. (1991) also reported results from a novel sentence-by-sentence listening task which provided another on-line measure of the time course of speech comprehension. In this task, participants simply listen to a series of individual sentences and press a button when they are ready to move on from one sentence to the next. After listening to all of the sentences, the listeners are then tested on their memory of particular words or propositions in the sequence of sentences they just heard. Ralston et al. used this task to study the time it took listeners to comprehend sentences produced either by a human voice or by the Votrax synthesizer. Figure 6 shows the average amount of time it took listeners to move from one sentence to the next in this task, for both fourth grade- and college-level passages presented in synthetic and natural speech. Ralston et al. found that listeners in this experiment took significantly longer to complete this task when the sentences were produced by the Votrax system than when they were produced by a human voice. Since the listeners in this experiment had also taken the MRT, Ralston et al. looked for correlations between segmental intelligibility scores and the differences in the sentence-by-sentence listening times for the two voices. They found that the measures from the tasks were significantly correlated with one another, suggesting that the time course of comprehension depends on the segmental intelligibility of the speech. However, since the r values for the correlations between the listening times and the MRT scores varied between +.4 and +.6, this analysis showed that segmental intelligibility could not account completely for the corresponding differences observed in comprehension between synthetic and natural speech.

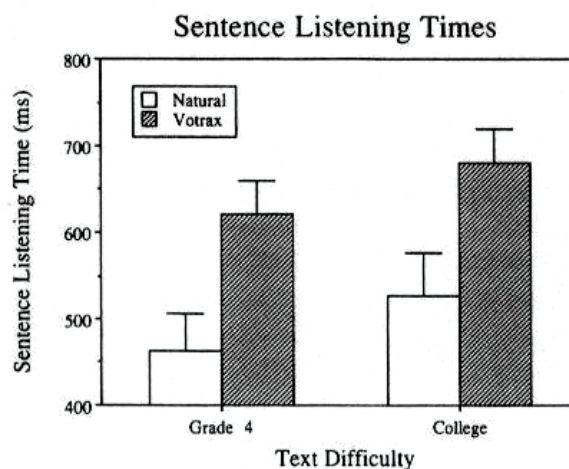


Figure 6. Average sentence-by-sentence listening times, by voice type and text difficulty. Error bars represent one standard error of the sample means (adapted from Ralston et al., 1991).

Paris, Gilson, Thomas and Silver (1995) followed up on Ralston et al.'s earlier study by investigating the comprehension of highly intelligible DECTalk speech, Votrax and natural speech. These investigators had listeners attend passively to passages produced by the three different voices, and then presented them with a series of true/false questions about the words and propositions which might have appeared in the various passages. In a separate condition, listeners were also asked to shadow (i.e., immediately repeat what they heard) two passages of synthetic or natural speech. Recall was better for both words and propositions produced by both DECTalk and natural voices than it was for items produced by Votrax. There were, however, no significant differences between either word or proposition recognition for DECTalk and natural speech. Nonetheless, shadowing accuracy was worse for DECTalk than it was for natural speech, and it was even worse for Votrax than it was for either of the other two voices. This on-line measure of perceptual processing indicated, once again, that there are perceptual difficulties in interpreting even the highest quality synthetic speech which disappear by the time the entire comprehension process has run its course.

Summary. Sensitive psycholinguistic tests of the time course of language comprehension have shown that it is more difficult to comprehend synthetic speech than natural speech. It takes longer, for instance, for listeners to verify whether or not synthetic sentences are true in a SVT than it does for them to verify naturally produced sentences. This effect holds even when natural and synthetic sentences are matched in terms of their segmental intelligibility, indicating that it is more difficult for listeners to process synthetic speech beyond the level of acoustic-phonetic interpretation than it is to process natural speech at this level. Processing deficits for synthetic speech also emerge in sentence-by-sentence listening tasks and shadowing tasks. Tests of the post-perceptual products of language comprehension have not always revealed greater listener difficulty in the comprehension of synthetic speech; however, listener recall of propositions from synthetically produced passages is often worse than recall of similar propositions from naturally produced passages. These findings suggest that listeners may be able to make up for the inherent difficulty of comprehending synthetic speech by implementing alternate processing strategies and drawing upon other sources of knowledge in order to complete a challenging comprehension task.

Point 5: Perception of Synthetic Speech Improves with Experience

Despite the fact that listeners consistently have more difficulty perceiving synthetic speech than natural speech, their ability to perceive synthetic speech typically improves if they simply receive more exposure to it. Such perceptual learning of synthetic speech has been documented for a wide variety of intelligibility and comprehension tasks.

Improvement on Broad Measures of Performance. In their initial studies of the perception of synthetic speech produced by the Haskins parallel formant resonance synthesizer, Nye and Gaitenby (1973) found that listener performance on the MRT improved significantly over the course of their experiment. In the first of six testing sessions—each of which contained 150 different test items—listeners averaged 17.5% errors when they heard synthetic speech; by the sixth (and last) of these sessions, however, they averaged only 8% errors. Listener performance on natural speech tokens in the same task, on the other hand, maintained an average error rate of about 4 to 5% throughout the six different testing sessions, but it was close to ceiling.

Carlson, Granstrom and Larsson (1976) tested blind listeners' ability to repeat sentences of synthetic speech and found that their performance on this task also improved dramatically between the first and the last testing sessions in the experiment. Carlson et al.'s listeners took part in eight separate testing sessions, the final four of which took place one week after the first four. In each of these testing sessions, the listeners heard a unique list of 25 sentences—each of which they had to repeat—at two different speaking rates. They also listened to a 7-minute long short story. Carlson et al. tabulated the number of words and sentences the listeners repeated correctly and found that average percent correct scores increased from 52% words correct and 35% sentences correct in the first session to 90% words correct and 77% sentences correct in the final session. Furthermore, Carlson et al. found that the listeners maintained their improved performance over the week-long break between the fourth and the fifth sessions. In the fourth testing session, listeners averaged 85% words correct and 70% sentences correct; one week later, in the fifth testing session, listeners averaged 83% words correct and 65% sentences correct. This pattern of results showed that significant improvements in the perception of synthetic speech could not only be made with relatively little exposure to synthetic speech (i.e., 100 sentences), but also that such improvements could be maintained over comparatively long-term intervals.

Rounsefell, Zucker and Roberts (1993) reported an even more dramatic demonstration of the speed at which exposure to synthetic speech improves listeners' ability to perceive it. They tested the ability of high school-aged listeners to transcribe a pair of sentences as produced by either the DECTalk, VoicaLite or Echo II synthesizers. Half of the listeners received training on the particular synthetic voice they were to hear before the testing sessions began. This training consisted of three repetitions of three different sentences, as produced by the synthetic voice, each of which was followed by a live repetition of the same sentence, produced by the experimenter in a natural voice. Rounsefell et al. found that listeners who were trained in this way—essentially hearing nine tokens of synthetic sentences before testing began—were significantly better than untrained listeners at a subsequent sentence transcription task. The trained listeners correctly transcribed, on average, 9.11 syllables out of the 14 syllables in the two test sentences; the untrained listeners, on the other hand averaged only 3.75 syllables correct.

Venkatagiri (1994) observed that, even though listeners' ability to perceive synthetic speech may improve rapidly after a few initial exposures, this perceptual improvement may not necessarily continue indefinitely. Venkatagiri asked listeners to transcribe sentences produced by the Echo II synthesizer, and investigated how much their transcription accuracy improved over three consecutive days of testing. On each of these three days, the listeners transcribed a series of 20 different synthetic sentences. Venkatagiri found that listeners' transcription performance improved significantly between days one and two of

testing (i.e., between the first 20 and the second 20 sentences), but that listener transcriptions were not significantly better on day three than they were on day two. Venkatagiri concluded that this failure to improve on day three may have been due to a ceiling effect, since the average percentage of correct transcriptions had already improved from 78.2% to 94.1% between days one and two, but were only able inch up to 96% on day three. Listeners' failure to post significantly higher scores on the third day of testing may therefore have been due to the fact that they had very little room left to improve beyond 94.1% correct. A task that does not set such easily reachable upper bounds on positive performance—as the sentence transcription task does—may yield continued perceptual improvements from exposure to synthetic speech over a longer interval of training.

Improved Reaction Times. As reviewed earlier in point 2, Slowiaczek and Pisoni (1982) found that listeners exhibit similar continued improvement in the perception of synthetic speech when tested on an auditory lexical decision task. They had listeners perform a lexical decision task, while listening to both natural and synthetic speech stimuli, over the course of five consecutive days. Figure 7 shows the average response time and the percentage of errors listeners made in this task, for each of the five consecutive days of training, for both synthetic and natural stimuli. This figure shows that listener response times for natural and synthetic non-word stimuli decreased significantly over the five days of training; however, the corresponding proportions of correct responses remained essentially constant for the duration of the study. This finding suggested that response time measurements might reflect continued improvements in perception even though performance on a coarser-grained measure such as response accuracy had already reached ceiling.

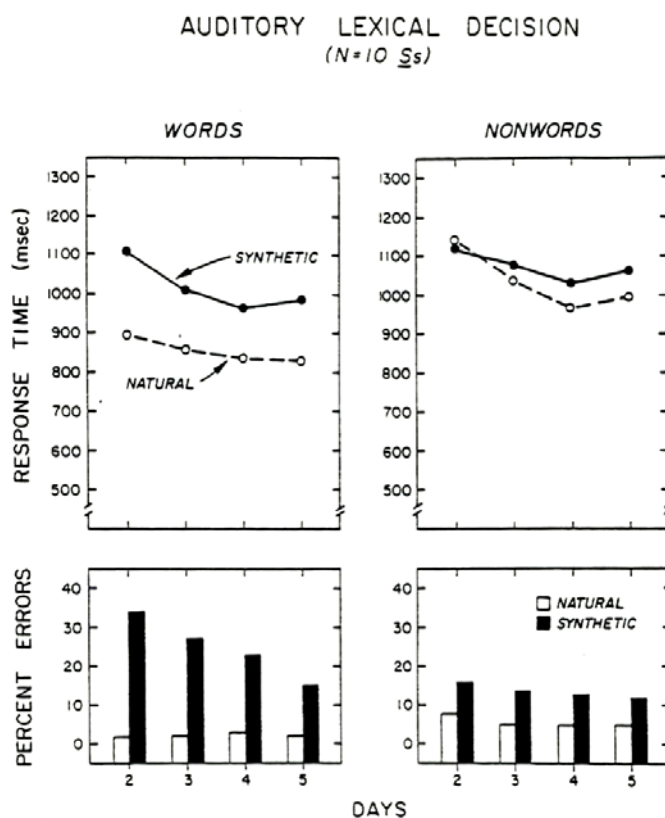


Figure 7. Lexical decision response times and errors, for words and non-words, by voice type and day of study (adapted from Slowiaczek & Pisoni, 1982).

Reynolds, Isaacs-Duvall, Sheward and Rotter (2000) also found that increased exposure to synthetic speech significantly decreased listeners' response times to synthetic speech sentences in a SVT. Reynolds et al. found that listeners exhibited such improvements even without explicit training on the SVT itself. One group of listeners in Reynolds et al. was trained for eight consecutive days—in between a sentence verification pre-test and post-test—on a sentence transcription task, in which they heard the same kinds of sentence stimuli that were used in the pre- and post-tests (as produced by DECTalk) without actually practicing the SVT itself. Another group of listeners in the study received no training or exposure to synthetic speech in between the sentence verification pre- and the post-tests. The authors found that trained listeners showed improvement between pre- and post-test in terms of the speed with which they verified synthetic false sentences; neither listening group's RTs decreased significantly for synthetic true sentences. Reynolds et al. suggested that this failure to improve may have resulted from a ceiling effect on performance with the true sentences, which had induced relatively short RTs in the initial pre-test. However, Reynolds et al.'s listeners also performed the SVT with naturally produced stimuli in both the pre- and the post-tests. Response times for these sentences were consistently shorter in both testing sessions than they were for the synthetically produced sentences. Moreover, Reynolds et al. found that the RTs for both groups of listeners decreased for these natural stimuli between the pre- and the post-test. Reynolds et al. suggested that this improvement may have resulted from increasing listener familiarity with the individual natural voices used to produce the stimuli.

Reynolds, Isaacs-Duvall and Haddox (2002) pursued the issue of listener improvement in the SVT further, questioning whether listeners' ability to comprehend synthetic speech could become as good as their ability to comprehend natural speech stimuli, after extended exposure to synthetic speech. Reynolds et al. thus had listeners participate in a five-day long testing sequence, during which they performed the SVT for 80 different sentences (40 produced by a natural voice, 40 produced by DECTalk) on each day. Reynolds et al. found that RTs were shorter for natural sentences than for synthetic sentences on all five days. Listener RTs also decreased significantly for both types of sentences between the first and the last days of the experiment. However, most of the improvement occurred between the first and second days of the experiment, and listener RTs had more or less bottomed out for both voices by the fifth day of testing. Nonetheless, Reynolds et al. charted regression lines to track the idealized future course of RT progress, beyond the fifth day of testing, for both the synthetic and the natural voices in this study. These regression lines indicated that the slopes of the natural and synthetic RTs were still diverging by the fifth day of testing. Even though listener performance improved for both voices in this study, that is, continued listener improvement on the natural voice was greater than continued improvement on the synthetic voice. This finding indicated that, no matter how much exposure listeners got to synthetic speech, their ability to comprehend it would still remain worse than their ability to comprehend natural speech, given equivalent amount of practice listening to both types of voices in a particular testing paradigm.

Pisoni and Hunnicutt (1980) also reported that listeners could rapidly improve their ability to comprehend synthetic speech over the course of an experiment. The listeners in this study answered multiple-choice questions based on passages they had either read, heard spoken by a natural voice, or heard spoken by the MITalk synthesizer. Table 3 shows the percentage of questions the listeners answered correctly in each half of the experiment, for all three presentation conditions. The percentage of questions that the listeners answered correctly increased between the first and second halves of the experiment for both types of auditory presentation, although the amount of improvement was greater when listeners heard the MITalk voice.

Although Pisoni and Hunnicutt did not run any tests to determine if the levels of improvement across both halves of the study were statistically significant, their results were nonetheless noteworthy because the comprehension of synthetic speech appeared to improve more than the comprehension of

natural speech—even though comprehension for both types of spoken passages was roughly equivalent during the first half of the experiment.

	1st Half	2nd Half
MITalk	64.1	74.8
Natural	65.6	68.5
Reading	76.1	77.2

Table 3. Percentage of comprehension questions answered correctly, by voice type and experiment half (adapted from Pisoni & Hunnicutt, 1980).

Perceptual Learning interacts with task type. Schwab, Nusbaum and Pisoni (1985) reviewed the improvements reported in the earlier studies of Pisoni and Hunnicutt (1980) and Slowiaczek and Pisoni (1982) and questioned whether or not the findings might be the result of listeners simply becoming more proficient at carrying out the experimental tasks, rather than becoming better at perceiving synthetic speech. In order to test this hypothesis, Schwab et al. gave listeners extensive training on a variety of perceptual tasks and investigated whether the type of stimuli used in the training had any effect on the amount of improvement the listeners made in these tasks. The tasks included: the closed-set MRT, word recognition using phonetically balanced (PB) word lists, transcription of (meaningful) Harvard sentences, transcription of (meaningless) Haskins sentences, and yes/no comprehension questions based on a series of prose passages the listeners had heard. On the first day of testing, listeners participated in all of these tests, in which they were played stimuli produced by the Votrax synthesizer. On days two through nine of the study, the listeners were split into two groups—one which received training on all of these various tasks and another which received no training at all. The group of listeners that received training on the tasks was further split into two sub-groups: those who heard synthetic (Votrax) speech stimuli during the training tasks, and those who heard natural speech stimuli during the tasks. After eight days of receiving training—or no training—all groups of listeners repeated the same battery of tests as on the first day of the experiment, in which they once again heard all stimuli as produced by the Votrax synthesizer.

Schwab et al. (1985) found that those listeners who had been trained on the various perceptual tasks with synthetic speech stimuli showed significantly greater improvement on the word recognition and sentence transcription tasks between pre- and post-test than did either the untrained group or the listeners who had been trained with natural speech stimuli. Furthermore, the natural speech group did not show any more improvement than the untrained control group. Schwab et al. interpreted these results as evidence that listener improvement in perceiving synthetic speech was not merely due to a practice effect, since the group trained with natural speech stimuli had received the same amount of training and practice with the tasks as the synthetic speech group, and yet they showed less improvement in perceptual performance between the pre-test and the post-test. Schwab et al. argued that the observed perceptual improvement was due to exposure to synthetic speech *per se*—that is, listeners became better at extracting acoustic-phonetic information from the synthetic speech signal as they gained increasing amounts of exposure to it. This perceptual ability is evidently domain-specific and cannot be acquired by simply being exposed to equivalent amounts of natural speech.

Schwab et al. (1985) reported that their listeners made the largest amount of improvement in the PB word recognition and meaningless sentence transcription tasks. The authors pointed out that these tasks had less constrained response sets than did the closed-set MRT and the meaningful sentence transcription tasks. Since listeners could depend less on higher-level semantic and structural information to perform well in the PB and meaningless sentence tasks, their correspondingly greater improvement in

these conditions must have been due primarily to developing their proficiency at interpreting the lower-level, acoustic-phonetic details in the synthetic speech signal. Corresponding improvements at interpreting such information in the MRT and the Harvard sentence tasks may have been masked by the listeners' ability to rely on higher-level linguistic information to perform these tasks well. Schwab et al. also ran a follow-up study, six months after the original training experiment, which showed that the group trained on synthetic speech stimuli still maintained better levels of performance on the various perceptual tasks than the group of listeners who were trained on natural speech. Schwab et al. therefore concluded that the effects of training had long-term benefits for the perception of synthetic speech.

Learning Effects: Generalization. Another way to interpret the results of Schwab et al. (1985) is that the natural speech group acquired little ability to generalize from their training with natural speech stimuli to the test conditions with synthetic speech stimuli. In another training study, Greenspan, Nusbaum and Pisoni (1988) questioned whether there were similar limits on the ability to make generalizations of particular synthetic speech training stimuli. Greenspan et al. thus adapted the training paradigm of Schwab et al. to determine how well listeners could generalize from particular sets of training stimuli to improve their performance on tests using novel synthetic speech stimuli. Greenspan et al. divided their listeners into five different training groups: one which received training on a new set of individual, novel words on each of the four successive days of training; a second group, which was trained on the same set of repeated words on each of the four training days; a third group, which listened to sets of novel sentences on all four days of training; a fourth group, which listened to the same set of repeated sentences on all four days of training; and a fifth group, which received no training on the four days between the pre and post-tests. The pre- and post-tests consisted of the closed-set MRT, the PB open-set word recognition task, and the transcription of Harvard and Haskins sentences. Each group of listeners—aside from the listeners who underwent no training at all—thus received training which was directly relevant to only two of these four testing conditions. Training on individual words was likely to benefit listeners on the MRT and PB word tasks, while training on sentences was likely to benefit listeners on the Haskins or Harvard sentence transcription tasks. Greenspan et al.'s training groups were also further divided between those who listened to novel stimuli everyday and those who listened to repeated sets of stimuli in order to determine whether it was easier for listeners to generalize from training sets which had greater amounts of variability in them.

While Greenspan et al. (1988) found no effects of novel versus repeated sets of training stimuli on listener performance in the testing conditions, they did find asymmetric effects of training between sentence and word recognition tasks. The groups that received training on the sentence transcription tasks showed improved performance on all four tasks in the study. The groups that received only training on individual novel or repeated words, however, only showed improved performance between pre- and post-test on the MRT and the PB word tasks; their performance levels on the two sentence transcription tasks remained the same. Greenspan et al. suggested that the failure of individual word training to transfer to the sentence transcription task may have been the result of the listeners' inability to parse the word boundaries in a stream of synthetic speech. The transfer of sentence training to the word recognition tasks, on the other hand, may have resulted from listeners simply being exposed to a large number of synthetic words in a wide range of contexts during the four days of training.

Greenspan et al. (1988) further investigated the perceptual consequences of variability in training stimuli in a follow-up experiment. In their second study, they trained two different listener groups on the PB word recognition task. After a 50-word pretest, one of these groups received training and feedback on the task with 200 different, novel words, while the other group received training on only 20 repetitions of 10 words which they had already heard in the pre-test. In a post-test, both groups of listeners were presented with a set of 10 words from the pre-test, along with 40 novel word stimuli. Both groups showed improvement in their ability to perceive these novel words; however, the group which had heard more

diverse training stimuli showed more improvement than the other group. Greenspan et al. concluded that improvement in the perception of synthetic speech depends, to a large extent, on the amount of acoustic-phonetic variability in the synthetic speech samples that listeners are exposed to during training.

Summary. Listeners' ability to perceive synthetic speech improves as they receive more exposure to it. Such improvement has been shown in a wide variety of behavioral tasks, including word recognition, sentence transcription, lexical decision, and sentence verification. This improvement can occur quite rapidly, even manifesting itself after exposure to only a few sentences of synthetic speech. Such perceptual improvement may still persist for as long as six months—and maybe even longer—after initial exposure to synthetic speech. Research has also shown, however, that there are limitations on the amount of improvement that listeners can make in the perception of synthetic speech. Even after substantial amounts of training on synthetic speech stimuli in a sentence verification paradigm, for instance, listeners cannot verify synthetic sentences as quickly as they do natural sentences. Furthermore, improvement in the perception of synthetic speech depends to some extent on the type of training or exposure that listeners receive. Training on the transcription of synthetic sentences improves listeners' ability to identify individual synthetic words, for instance, but training on the identification of individual synthetic words does not improve listeners' ability to transcribe whole synthetic sentences. Other findings also indicate that the amount of improvement listeners make in the perception of synthetic speech depends on the amount of variability in the training stimuli they are exposed to. Increased familiarity with synthetic speech may thus alleviate, but not overcome, the fundamental limitations that poor acoustic-phonetic quality places on listeners' ability to perceive synthetic speech as well as they perceive natural speech.

Point 6: Alternative Populations Process Synthetic Speech Differently

The studies reviewed above have found consistent perceptual deficits in the processing of synthetic speech by listeners who are typically college-aged, normal-hearing, native speakers of English. Studies which have investigated how other groups of listeners perceive synthetic speech have typically found that these alternative populations of listeners process synthetic speech differently than their college-aged, normal-hearing, native-speaker counterparts. In most cases, these alternative groups have even more difficulty perceiving and processing synthetic speech.

Non-Native Listeners. Greene (1986), for instance, found that non-native listeners of English have significantly more difficulty perceiving synthetic speech (in English) than native listeners do. Greene tested both native and non-native speakers of English on the MRT and a sentence transcription task, using stimuli produced by a natural voice and MITalk. Table 4 shows the percentage of synthetic and natural words that both groups of listeners correctly identified in the MRT. The non-native listeners in this study performed only slightly worse on this task for the natural voice than the native listeners did, with scores at or near ceiling. However, the performance gap between the two listening groups significantly increased when they listened to MITalk speech, with the native listeners remaining near ceiling but the non-native listeners decreasing substantially in accuracy.

	Natural	MITalk
Natives	99	93
Non-Natives	95	86

Table 4. Percent Correct on MRT by voice type and listener group (adapted from Greene, 1986).

However, Greene (1986) found no interaction between listening group and voice type in the transcription of either semantically anomalous or meaningful sentences. The non-natives' percent correct sentence transcription scores were consistently worse than the natives' scores by the same amount, for both natural and MITalk voices, as shown in Tables 5a and 5b below. This may be due to the fact that the native listeners performed at ceiling for the natural voice for both sentence types, and near ceiling for the MITalk voice for meaningful sentences.

	Natural	MITalk
Natives	98	79
Non-Natives	70	51

Table 5a. Words correctly transcribed from semantically anomalous sentences, by voice type and listener group (adapted from Greene, 1986).

	Natural	MITalk
Natives	99	93
Non-Natives	70	64

Table 5b. Words correctly transcribed from meaningful sentences, by voice type and listener group (adapted from Greene, 1986).

Despite the lack of an interaction in the sentence transcription scores, the gap between native and non-native listeners' scores was much greater in the sentence transcription task than it was in the MRT. This result indicated that native listeners were significantly better than non-natives in drawing upon higher-level linguistic knowledge to interpret whole sentences produced in synthetic speech—along with being more proficient at interpreting the low-level acoustic cues for individual synthetic speech segments in isolated words.

Greene (1986) also noted that non-native listeners not only scored worse on both the MRT and the sentence transcription tasks, but they also showed a wider range of variability in their percent correct scores than did their native-listener counterparts. Greene tested another group of non-native listeners on a sentence transcription task and found that their percent correct scores on this test correlated highly with their scores on the TOEFL ($r = +.83$) and on the English Proficiency Test ($r = +.89$). Greene thus concluded that the ability of non-native listeners to perceive synthetic speech depended greatly on their proficiency in the language being spoken.

More recently, Reynolds, Bond and Fucci (1996) also found that non-native listeners transcribed synthetic speech sentences less accurately than native English listeners. The participants in their study were asked to listen to pairs of thematically related sentences and then transcribe the second sentence in each pair. Reynolds et al. presented these sentence pairs to both native and non-native listeners of English in quiet and noisy (multi-talker babble at +10 dB SNR) conditions. The authors found that transcription accuracy was significantly higher for native listeners (96.3% correct) than non-native listeners (54.5% correct) in the quiet condition. The authors also found that introducing babble noise into the signal had a significantly greater detrimental effect on the non-natives' performance; their percent correct scores dropped 8.7% between quiet and noisy conditions, while the native listeners' scores dropped only 2.8%. Reynolds et al. suggested that this drop in performance probably occurred simply because of the non-native listeners' relative unfamiliarity with the English language and their corresponding inability to

interpret unusual new forms of it. Furthermore, Reynolds et al. noted a wide variation in percent correct transcription scores in their study and suggested that, as in Greene (1986), non-natives' ability to perceive synthetic speech depended in large part on their proficiency in the English language.

Children. Children form another group of “alternative listeners.” Greene and Pisoni (1988) studied the perception of synthetic speech by children and found that they, like their college-aged counterparts, had more difficulty processing synthetic speech than natural speech, on a variety of perceptual tasks. Greene and Pisoni developed a four-alternative, forced-choice, picture-word matching task, using items from the Peabody Picture Vocabulary Test, to investigate children's ability to understand synthetic speech. Kindergartners who participated in this test correctly identified 82% of words produced by the Prose 2000 speech synthesis system (Groner, Bernstein, Ingber, Pearlman & Toal, 1982). They correctly identified 94% of natural speech tokens, however. Second graders who took this test performed better, with 94% correct for synthetic speech tokens and 98% correct for natural speech tokens. Therefore, the performance of both groups of children on this task was analogous to the levels of performance obtained with adult listeners, in that they tended to perceive synthetic speech less well than natural speech.

Greene (1983) found that children show similar deficits in their ability to recall synthetic speech items. Greene tested the ability of fourth graders to recall in any order the items on lists of two to eight digits in length, as produced by either a natural or a synthetic voice. The children in this task correctly recalled 82.9% of the items in the naturally produced lists, but only 78.5% of the synthetically produced list items. Greene got similar results from a task involving the free recall of nouns by children. In this task, children listened to strings of nouns, spoken in either a natural or a synthetic voice, and were asked to repeat back the items in the string in any order. All of the strings were from two to 10 words in length. Children consistently recalled more words correctly from the naturally produced lists than they did from the synthetically produced lists. They also had more difficulty, for both types of voices, in recalling items from the longer lists. This pattern of results replicated earlier findings from studies using adult listeners such as Luce et al. (1983), which showed that the recall of synthetically produced items is more difficult than the recall of naturally produced items for college-aged listeners, as well, due to the comparative difficulty they have in interpreting and encoding the acoustic-phonetic structure of synthetic speech.

Mirenda and Beukelman (1987) directly compared the performance of adults and children on Greene's (1983) battery of perceptual tasks, using natural and synthetic stimuli. Table 6 shows the percentages of correct responses adults and children gave in transcribing individual words presented in four different voices (natural speech, DECTalk, Echo and Votrax synthesizers). These results showed that children consistently performed worse on this single-word transcription task than adults, for both natural and synthetic voices. Also, the younger children listeners—who were between 6 and 8 years old—performed worse than the older children, who were between 10 and 12 years old.

	Natural	DECTalk	Echo	Votrax
Adults	99.2	78.0	39.6	57.2
10-12	96.9	76.0	37.6	56.4
6-8	93.6	72.0	34.0	48.4

Table 6. Percent individual words correctly transcribed, by voice type and listener group (adapted from Mirenda & Beukelman, 1987).

Since children performed worse than adults in this task for all voices, the greater difficulties that children seemed to have in perceiving synthetic speech may, in fact, have been due to the greater difficulties they had in performing the perceptual task itself. Miranda and Beukelman (1987) also administered a sentence transcription task to these child and adult listeners; Table 7 shows the percentage of words in these sentences that each group of listeners transcribed correctly, for each of the four voices. These results show an even greater discrepancy between the performance of adults and listeners in this task than there was in the individual word transcription task.

	Natural	DECTalk	Echo	Votrax
Adults	99.3	96.7	68.2	83.8
10-12	96.4	90.2	61.8	68.2
6-8	94.2	81.1	35.8	57.1

Table 7. Percent words correctly transcribed from meaningful sentences, by voice type and listener group (adapted from Miranda & Beukelman, 1987).

All groups performed better on this task than they did on the individual word transcription task—especially for the synthetic voices. This result indicated that children, like adults, draw on higher-level linguistic knowledge to interpret synthetic speech. Since the adult transcription scores improved in the sentence transcription task more than the children’s did, however, this pattern of results suggested that the adult listeners were better at accessing and using this higher-level linguistic information than the children were.

Children’s Comprehension of Synthetic Speech. Reynolds and Fucci (1998) reported that children also have more difficulty comprehending synthetic speech than natural speech. They measured children’s comprehension of naturally and synthetically (DECTalk) produced stimuli using a SVT. The children they tested, who were between the ages of 6 and 11, consistently responded faster to natural sentences than synthetic sentences. Reynolds and Jefferson (1999) expanded upon this finding by testing child listeners from two separate age groups, 6 to 7 years old and 9 to 11 years old, in an identical experimental paradigm. They found an equivalent processing advantage for natural versus synthetic sentences for all listeners, but also found that the 9 to 11 year-olds comprehended both kinds of sentences faster than the 6 to 7 year-olds. This pattern of results mirrored the findings of Miranda and Beukelman (1987; 1990) for similar groups of listeners performing transcription tasks. It is thus possible that the differences in sentence verification times between 6-7 year-old and 9-11 year-old listeners may be due to the corresponding difficulties each group has in interpreting the acoustic-phonetic structure of synthetic speech, rather than higher-level comprehension difficulties in post-perceptual processing. However, the differences between the two groups may also reflect the extent to which older children are simply better at doing perceptual tasks of this kind.

Reynolds and Jefferson (1999) also tested their listeners again in a second session, either 3 or 7 days after the first day, on the same task. They found that the 6 to 7 year-olds had significantly faster response times in the second testing session than they did in the first, but that they 9 to 11 year-olds showed no significant differences in response times between sessions. The younger children’s RTs improved primarily in their responses to synthetic sentences. Reynolds and Jefferson suggested that such improvement may have been due to the 6 to 7 year-old group’s perceptual systems being more flexible than those of the older children, and thus more capable of quickly adjusting to novel stimuli. However, the 6 to 7 year-olds’ RTs were quite long in the first session—especially for synthetic sentences—and

thus left ample room for improvement, while the 9 to 11 year-olds' RTs may have initially been closer to ceiling performance.

Older Adults. Little is known about the extent to which older adults' perception of synthetic speech differs from that of younger adults. Sutton, King, Hux and Beukelman (1995) investigated perceptual differences between listener preferences for particular rates of synthetic speech. Two groups of listeners were used in this experiment: 20 older adults from 61 to 79 years of age and 20 younger adults from 21 to 28 years of age. All listeners heard sentences produced by DECTalk, which ranged in rate from 120 to 250 words per minute. After hearing each sentence, the listeners indicated to the experimenter what their subjective comfort level was in hearing the sentence, on a Likert scale from 1 to 7; a rating of 1 indicated "too slow" while a rating of 7 indicated "too fast." Sutton et al. found that younger adults had a broad "comfort range"—which was defined as sentences receiving scores between 3 and 5 on the Likert scale—for sentences spoken at rates between 150 to 220 words per minute. The older listeners, on the other hand, preferred slower rates of synthetic speech, from 130 to 210 words per minute. Sutton et al. suggested that a number of different factors might lead older listeners to prefer slower rates of synthetic speech—among them a decreased ability in temporal processing, hearing loss, and global changes in auditory perception.

Hearing-Impaired Listeners. Studies investigating older listeners' perception of synthetic speech have often been framed within the context of hearing impairment and its potential effects on the perception of synthetic speech. Kangas and Allen (1990), for instance, tested two groups of listeners between the ages of 49 to 64; the members of one group had normal hearing while the members of the other group had all acquired hearing losses. Both groups of listeners transcribed individual words that were produced by either DECTalk or a natural male voice. Kangas and Allen found that both groups of listeners transcribed naturally produced words correctly more often than synthetically produced words; they also found that hearing-impaired listeners performed worse on this task than normal-hearing listeners did. Importantly, there was no interaction between voice type and listener group; this suggested that synthetic speech did not exacerbate the difficulty these hearing-impaired listeners had in perceiving speech. Rather, the performance of hearing-impaired listeners in this task could best be understood by simply combining the deficits that synthetic speech and hearing impairment both impose individually on normal speech perception processes. Interestingly, Kangas and Allen also noted that there was much more individual variability in the hearing-impaired listeners' synthetic speech transcription scores than there was in the normal-hearing listeners' scores. Kangas and Allen concluded that most of this variability could be accounted for by the corresponding variability in the hearing-impaired listeners' performance on the natural speech transcription task.

In another study of hearing-impaired listeners, Humes, Nelson and Pisoni (1991) found that age also does not interact with hearing-impaired listeners' performance on the MRT. They tested a group of 66-73 year-old hearing-impaired listeners along with two groups of 19-22 year-old listeners on both the closed and open-response formats of the MRT. One of the groups of 19-22 year-old listeners heard the MRT items in the clear, while the other heard them through spectrally-shaped noise, which was designed to mimic the effects of hearing loss. Humes et al. included this condition in order to test the extent to which the decreased performance of the hearing-impaired listeners—in comparison to the younger listeners—was due to their hearing-impairment, rather than their increased age. All groups of listeners heard the MRT items as they were produced by a natural voice and by both the DECTalk and Votrax synthesizers. Figure 8 shows the percentage of word recognition errors each group of listeners made on the MRT while listening to the three different voices. All listeners in Humes et al. performed better on the open-response MRT when they heard the DECTalk and natural voices than when they heard the Votrax synthesizer. There were no significant differences between performance levels on DECTalk and the natural voice, except for the normal-hearing young listeners, who showed a small but significant

advantage with the natural voice. These normal-hearing listeners were also better at the MRT task, for all three voices, than either the hearing-impaired or masked-noise listening groups. There were no significant differences between the hearing-impaired and masked-noise groups, indicating that the decreased level of performance on the MRT by the older, hearing-impaired listeners could be accounted for solely by their hearing impairment, regardless of their age. However, greater variability was observed within the group of hearing-impaired listeners than for either of the younger groups of listeners. Humes et al. pointed out that individual performance on the MRT by the hearing-impaired listeners corresponded closely to their level of hearing loss in pure-tone average ($r = -.73, -.75, -.8$ for the natural, DECTalk, and Votrax voices, respectively). Their performance in the synthetic speech condition was also strongly correlated with their performance on the natural speech condition ($r = +.96$ and $+.90$ for DECTalk and Votrax, respectively). As in the earlier study by Kangas and Allen (1990), therefore, hearing impairment did not interact with the degraded quality of the synthetic speech to produce further perceptual difficulties for the listener under the MRT testing conditions.

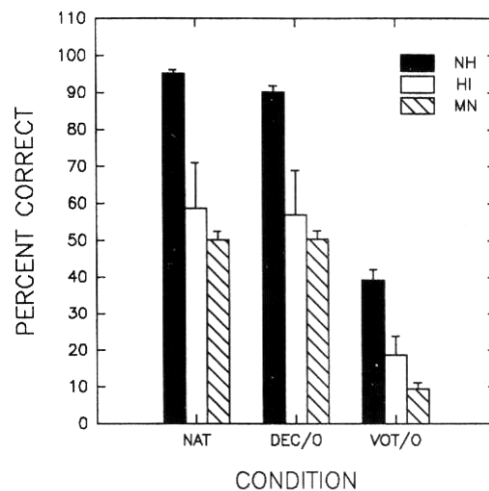


Figure 8. Percent of words correctly identified on the MRT, by voice type and listener group (NH = normal-hearing; HI = hearing-impaired; MN = masked-noise) (adapted from Humes et al., 1991).

Humes, Nelson, Pisoni and Lively (1993) tested older listeners' perception of synthetic speech using a serial recall task. The listeners in this study included both younger (21-24 years old) and older (61-76 years old) normal-hearing adults, who were presented with sequences of 10 words and were asked to repeat those words in the order that they heard them. The listeners heard these lists in both natural and synthetic voices. After performing the serial recall task, Humes et al. also had listeners transcribe all of the individual words that had been used in the lists (both synthetic and natural) in order to measure the intelligibility of the test words. The results of these two tests replicated earlier findings that synthetic speech was both less intelligible and more difficult to recall than natural speech. While the younger listeners had a slight advantage in the recall task, there were no significant differences between age groups on the transcription task, indicating that the processing of the acoustic-phonetic structure of synthetic speech does not significantly deteriorate with age in normal-hearing listeners.

Expert Listeners. One final group of alternative listeners—which the existing literature on synthetic speech perception has hardly studied—is expert listeners. With the increasing prevalence of speech synthesis products on the market today, this population of listeners is undoubtedly growing. Blind

listeners, for example, may benefit from speech synthesis programs on their computers which can output text for them to hear in the form of speech. Such listeners have extensive practice listening to the specific form of synthetic speech that their computers produce. Their ability to extract information from such synthetic speech may be far more advanced than those that listeners in the short-term training experiments, such as Schwab et al. (1985), may be able to develop over a 10-day period.

While no studies that we know of have directly investigated the perceptual abilities of such “expert listeners,” Hustad et al. (1998) did compare the perceptual abilities of a group of “speech synthesis experts” to those of a group of speech-language pathologists (who had never had more than incidental contact with synthetic speech) and another group of listeners who were completely unfamiliar with synthetic speech. Hustad et al.’s “speech synthesis experts” all had extensive experience listening to synthetic speech—they all had worked on a regular basis with both DECTalk and MacinTalk Pro (the two synthetic speech voices used in the study) for at least five hours a week for a year or more prior to the study. All listeners in this study took the MRT, listening to items produced by both DECTalk and MacinTalk Pro voices. Not surprisingly, the expert listeners outperformed both the speech-language pathologists and the inexperienced listeners on this task, in both voice conditions. Hustad et al. thus concluded that extensive experience listening to synthetic speech improved listeners’ ability to extract information from the synthetic speech signal. Moreover, Hustad et al. reasoned that similar amounts of experience listening to natural speech in an analytic fashion could not improve the perception of synthetic speech, since the speech-language pathologists did not outperform the inexperienced listeners on the MRT. This finding is analogous to the results of Schwab et al. (1985), who found that only training on synthetic speech—and not natural speech—improved listener performance on recognizing words and transcribing sentences produced in synthetic speech. The extent to which expert listeners’ performance on the same task might differ from performance by listeners who have only been trained on synthetic speech stimuli for a period of eight days, however, remains unknown.

Summary. The research on “alternative” groups of listeners indicates that these listeners display more variation in perceiving synthetic speech than is commonly revealed in experiments on the perception of synthetic speech by young, normal-hearing, native listeners. Non-native listeners and children, for instance, have more difficulty perceiving synthetic speech than native-speaking, college-aged listeners in a variety of behavioral tasks, including the MRT, sentence transcription, lexical recall and sentence verification. “Expert” listeners, on the other hand, appear to have developed an ability to extract information from synthetic speech signals which surpasses those which inexperienced listeners are able to develop in laboratory training experiments. However, the perceptual abilities of highly experienced listeners remain largely unstudied, along with those of older adults. Most research on older adults’ perception of synthetic speech has focused on the effects of hearing impairment on the perception of synthetic speech; this research has shown that hearing impairment does not interact with the poor quality of synthetic speech to further degrade synthetic speech intelligibility. Accounting for this wide range of variability in listeners’ perception of synthetic speech is an important consideration in the development of robust speech synthesis applications.

Point 7: Prosodic Cues, Naturalness and Acceptability.

Research on the perception of prosodic or suprasegmental features in synthetic speech has often approached prosodic quality as a matter of subjective listener preference—a factor that could possibly make certain kinds of synthetic speech more acceptable, or natural-sounding, to listeners. Researchers have thus treated prosody primarily as a voice quality issue to be dealt with once the initial research goal of making synthetic speech as intelligible and comprehensible as natural speech had been met. The fact that synthetic speech may often sound monotonous or unnatural to listeners might at that point be investigated independently to improve highly intelligible speech synthesis for the sake of listener comfort

and preference. This approach to researching prosody and naturalness in synthetic speech treats naturalness and prosody as independent of the intelligibility or comprehensibility of the synthetic speech signal (Nusbaum, Francis & Henly, 1995). It contrasts sharply with another line of research which has shown that the intelligibility of synthetic speech depends to some extent on the quality and appropriateness of the prosodic cues in the speech signal (Slowiaczek & Nusbaum, 1985; Paris, Thomas, Gilson & Kincaid, 2000; Sanderman & Collier, 1997).

Prosodic Influences on Subjective Judgments of Naturalness, Preference, and Acceptability.

One early study on the subjective evaluation of synthetic speech was carried out by Nusbaum, Schwab and Pisoni (1984), who investigated subjective responses to synthetic and natural speech passages by simply asking listeners how they would rate the voices in the passages on a variety of subjective impressionistic scales (e.g., interesting vs. boring, grating vs. melodious, etc.). Nusbaum et al. also asked listeners to indicate how much they would trust each voice to provide them with particular kinds of information (e.g., tornado warnings, sports scores, etc.). Finally, Nusbaum et al. presented their listeners with a series of comprehension and recall questions based on the passages they had heard produced by the three different voices used in the study (natural speech, MITalk, and Votrax) and then asked the listeners how confident they were that they had comprehended the passages correctly.

The authors found that the listeners tended to give the natural voice higher ratings than both synthetic voices on descriptions that would be considered preferable (e.g., smooth, friendly, polished). A few adjectives (easy, clear, pleasant and fluent) also teased apart listener preference for the two different synthesizers in that MITalk (which was more intelligible than Votrax) was also judged more preferable by the listeners in these descriptive terms. Listeners also placed the most trust in the natural voice, followed by MITalk and then Votrax. Although listeners' comprehension of passages produced by all three voices was essentially equivalent, regardless of the voice used, their confidence in their ability to comprehend these passages was much lower for the two synthetic voices. This finding indicated that the post-perceptual comprehension task—which essentially asked listeners to recall particular facts and words, or to draw inferences based on the passages they had just heard—was not sensitive enough to capture subtler difficulties in the process of comprehending synthetic speech that the listeners were consciously aware of.

Terken and Lemeer (1988) assessed the interaction between naturalness and intelligibility in terms of their respective influences on the perceived “attractiveness” of individual Dutch sentences produced by speech synthesis. They recorded a passage of 21 sentences, read in a natural voice, and then re-analyzed the entire recorded passage using LPC synthesis. Terken and Lemeer did this twice, once using 30 LPC coefficients and another time using only 6; the re-analysis with 30 coefficients produced a version of the passage with relatively high segmental intelligibility, while the version with 6 LPC coefficients had low segmental intelligibility. Terken and Lemeer also created additional versions of these passages—for both LPC re-analyses—using either a copy of the natural prosodic contour or just a flat (monotone) pitch contour throughout. They then played all four versions of these passages—either whole or in a sequence of individual sentences—to a group of Dutch listeners. In each condition, the listeners were instructed to rate how “attractive” each stimulus sounded on a scale from 1 to 10. When listeners heard the entire passage of 21 sentences all at once, Terken and Lemeer found that both intelligibility and prosody affected the ratings of attractiveness: listeners rated highly intelligible passages as more attractive than the less intelligible passages, and they rated the versions with natural prosody as more attractive than those produced in monotone. Interestingly, the effect of segmental intelligibility on the attractiveness ratings was stronger than the effect of prosody.

The attractiveness judgments the listeners made when they heard the passages one sentence at a time yielded interesting interactions between the intelligibility and prosody factors. For the highly intelligible sentences, listeners judged the natural prosody versions to be more attractive than those

produced in monotone; for the low intelligibility sentences, however, listeners rated both prosodic versions as equally attractive. Terken and Lemeer suggested that these results could be accounted for by assuming that the perception of prosody is strongly dependent on segmental intelligibility. Listeners evidently need time to adjust to speech with poor segmental quality before they can make judgments about its prosodic qualities. While listening to synthetic speech one sentence at a time, listeners apparently do not have enough time to perceive prosodic differences in poorly intelligible speech; however, with highly intelligible speech, they do, and it is under these conditions that the prosodic contributions to perceived “attractiveness” emerge.

In more recent studies of prosody, researchers have looked at listeners’ subjective assessments as a means of evaluating the quality of particular prosodic features in synthetic speech. Terken (1993), for instance, evaluated systems of prosodic rules for synthetic speech by playing passages produced with those rules to “experienced phoneticians” and then asking them to rate the naturalness of each passage on a scale from 1 to 10. The listeners heard two versions of each passage—one which played the entire, 10 sentence-long passage in order, and another version which scrambled the order of the individual sentences within the passage. Both versions of these passages were produced by a diphone synthesizer using either the natural pitch contour or one of two different sets of synthetic prosody rules: an older set of prosodic rules calculated declination patterns over the entire utterance, while a newer set of rules established intonational boundaries within the utterance and reset the declination pattern after each one. Terken’s listeners rated the passages produced with the newer set of synthetic prosody rules as more natural-sounding than the passages produced with the older set of rules, in both the scrambled and ordered-passage conditions. In the scrambled passages condition, the new set of rules not only improved the perceived naturalness of individual sentences over the old set of rules, but also made the sentences sound as natural to the “expert phoneticians” as the utterances produced with a natural intonation pattern. In the complete, ordered text condition, however, the passages with natural intonation sounded significantly more natural to the listeners than the passages with either set of synthetic prosody rules. Furthermore, the naturalness of only the passages with the natural intonation pattern was greater in the ordered condition than in the scrambled condition; the naturalness of the passages with the synthetic rule sets was slightly worse in the ordered condition. This pattern of results indicated that there are discourse-level prosodic patterns in the naturally produced speech which make it sound more natural than texts which have been produced with only sentence-level prosodic rules.

Sanderman and Collier (1996) further developed Terken’s (1993) sets of prosodic rules and assessed how their improved rule sets affected listeners’ acceptance of and preference for synthetic speech. Their new sets of prosodic rules essentially induced varying levels of boundary strength into synthetically produced pitch contours by independently adjusting features for phrase contour, boundary pause length, and declination reset. Sanderman and Collier played pairs of identical synthetic sentences—one of which had these prosodic phrasing rules and one of which did not—to a group of listeners and asked them which sentence of the pair they preferred. A listener preference for the prosodically phrased sentences emerged for sentences which were longer than about nine words; for sentences that were shorter than this, the listeners indicated no clear preference. In a subsequent experiment, Sanderman and Collier tested listener acceptability of sentences that had been produced with prosodic rule sets using different numbers of possible phrase boundary strengths. One set of sentences had two possible phrase boundary strengths, another had three, another had five, and yet another set of sentences was produced with natural prosody. Sanderman and Collier played sentences produced with each set of prosodic rules to untrained listeners and then asked them to rate how acceptable each sentence was on a scale from 1 to 10. In general, the listeners preferred the sentences produced with natural prosody to those produced with the artificial prosody rules. There was no significant difference in acceptability between the sentences with natural prosody and those with five different possible phrase boundary strengths, however. Sanderman and Collier’s results therefore indicated that appropriate prosody not only becomes a more important

contributor to the naturalness of synthetic speech as the synthetic speech segment becomes longer, but also that greater amounts of variability in the implementation of prosodic rules in synthetic speech significantly improves the perceived naturalness of that speech.

Prosody and Naturalness Interact with Comprehension and Intelligibility. Early studies on the perception of prosody in synthetic speech yielded only marginal evidence that prosodic cues contributed significantly to the overall intelligibility of synthetic speech. Slowiaczek and Nusbaum (1985) investigated the importance of prosody to synthetic speech intelligibility by presenting listeners with both meaningful (Harvard) and meaningless (Haskins) sentences, produced with either a “hat-pattern” (typical of declarative sentences) or a flat (monotone) pitch contour. Listeners also heard the sentences at two different speaking rates—150 and 250 words per minute—as produced by the Prose 2000 system; their task was to simply transcribe the words as they heard them. Slowiaczek and Nusbaum found that both speaking rate and semantic content influenced the accuracy of listeners’ transcriptions. Percent correct transcriptions were significantly higher for sentences produced at a slower rate of speech and for meaningful sentences. Slowiaczek and Nusbaum did not, however, find that the “hat pattern” increased transcription accuracy over the monotone pitch contour. Appropriate prosody, that is, did not have a consistent effect on individual word intelligibility.

In a follow-up experiment, Slowiaczek and Nusbaum (1985) explored the possibility that appropriate prosody might improve intelligibility if listeners heard a wider variety of syntactic constructions in the sentences they had to transcribe. They expanded upon the paradigm they used in their first experiment by including declarative sentences with active, passive and center-embedded syntactic constructions. These various sentences could also be either long or short. Slowiaczek and Nusbaum found that the listeners had less success transcribing the longer sentences and the center-embedded sentences in this study. They also found that sentences with the “hat pattern” prosody proved easier for the listeners to transcribe than sentences with monotone pitch contours. Their findings suggest that prosodic information in synthetic speech is useful to a listener when the syntactic structure of a sentence is not predictable.

Nusbaum et al. (1995) maintained that subjective judgments of preference or acceptability were highly dependent on the intelligibility of the synthetic speech signal. Nusbaum et al. therefore attempted to develop measures of naturalness that were independent of segment intelligibility. For example, in one experiment, Nusbaum et al. attempted to measure the “naturalness” of synthetic glottal pulse sources. The authors constructed one second-long vowels (/i/, /u/ and /a/) from sets of either one or five individual glottal pulses excised from both natural and synthetic (DECTalk, Votrax) speech. Nusbaum et al. played these vowels to listeners and asked them to identify—as quickly as possible—whether the vowels had been produced by a human or a computer. The “naturalness” of any particular vowel stimulus in this task was taken to be the probability that listeners would classify it as having been produced by a human. This classification experiment yielded an unexpected pattern of results: the “naturalness” of any given stimulus depended on both its vowel quality and the particular voice with which it was produced. The listeners consistently identified the /u/ vowels, for instance, as having been produced by a computer. /a/ vowels, on the other hand, exhibited an unexpected pattern of identification: DECTalk /a/ was more consistently identified as “human” than either of the natural /a/ productions (which, in turn, were more natural than the Votrax /a/). Only /i/ vowels were identified by listeners along the classification hierarchy that Nusbaum et al. expected: natural productions were consistently identified as “human” more often than DECTalk /i/, which was also more consistently classified as “human” than Votrax /i/. Nusbaum et al. suggested that this mixed pattern of results may have emerged because the higher frequency components of the glottal source waveform—which are present in the second formant of /i/ but not of /u/ or /a/—are important to identifying its naturalness.

In another experiment, Nusbaum et al. (1995) attempted to assess the naturalness of prosodic information independently of the influences of segmental intelligibility. To accomplish this, they low-pass filtered words produced by two human talkers and two synthetic speech voices (DECTalk and Votrax) at 200 Hz, thus removing segmental cues but preserving the prosodic information in the original speech sample. Nusbaum et al. then played these low-pass filtered stimuli to listeners and asked them to determine as quickly as possible whether the stimuli had been produced by a human or a computer. Listeners identified these stimuli just as they identified the /i/ stimuli in the previous experiment: the human voices were consistently more “natural” than the DECTalk voice, which, in turn, was more “natural” than the Votrax voice. Nusbaum et al. thus concluded that low-pass filtered words provided a reliable, intelligibility-free measure of the naturalness of any given voice. The results of this study also indicated that the lexical-level prosody of even a high-quality synthesizer such as DECTalk was still clearly inferior to the prosody of natural speech.

Sanderman and Collier (1997) showed that generating synthetic sentences with appropriate prosodic contours facilitates comprehension—and that, moreover, inappropriate prosodic contours may make comprehension more difficult. They demonstrated the importance of prosody to synthetic speech comprehension by investigating listener interpretations of a series of syntactically ambiguous sentences. For example, one of the sentences in Sanderman and Collier’s study was (translated from the Dutch) “I reserved a room in the hotel with the fax.” This sentence is ambiguous because “with the fax” may describe either how the room was reserved or a particular facility that the hotel has. Prosodic phrasing may help disambiguate these two interpretations. For instance, placing an intonation break between “the hotel” and “with the fax” discourages listeners from grouping these two phrases together within the same noun phrase; hence, this intonational phrasing supports the interpretation wherein “with the fax” describes how the reservation was made. Without an intonation break in “the hotel with the fax,” listeners are more likely to interpret the fax as being one of the hotel’s facilities. However, without any disambiguating information in sentences like these, listeners tend to be biased towards one interpretation of the sentence over another. To measure this bias, Sanderman and Collier presented written versions of sentences like the one above to untrained listeners of Dutch and asked them to circle a paraphrase of the most likely interpretation of the sentence. Using this method, they determined what the most likely interpretation of each ambiguous sentence was: the most likely interpretation of the example sentence, for instance, was the one in which “with the fax” described how the hotel reservation had been made.

Sanderman and Collier (1997) investigated how prosodic phrasing might interact with the inherent interpretive bias in these syntactically ambiguous sentences by having listeners answer questions which might further bias them towards one interpretation over another. Listeners read a question such as “How did I reserve the hotel room?” and then answered it based on information they heard spoken to them in a synthetically produced sentence. The prosodic rule set used to produce these sentences enabled five different levels of phrase boundary strength—this rule set being the one that produced the most natural-sounding stimuli in Sanderman and Collier (1996). Each particular sentence was produced with one of three phrasings: one which supported the most likely interpretation of the sentence, another which supported the less likely interpretation of the sentence, and another which had no phrasing boundaries. This “zero” prosody version was included in order to establish a response baseline with which to compare the effects of the other two prosodic phrasings.) Sanderman and Collier (1997) measured the amount of time it took listeners to respond to the written question after they had heard the inquired-after information in the synthetically produced answer (e.g., the fax). Sanderman and Collier (1997) found facilitatory effects for matched questions and answers: a less likely answer combined with a less likely question reduced response times in comparison to the “zero” prosody answers, just as more likely answers in conjunction with more likely questions did. The mismatched conditions yielded a different set of effects: a more likely response matched to a less likely context question significantly increased response times over the baseline condition; however, less likely answers matched to more likely context questions did not. Thus, it appears

that listeners have more difficulty undoing a bias towards a less likely interpretation of the sentence. The broader implication of Sanderman and Collier's (1997) work is that appropriate prosodic phrasing is necessary for optimal comprehension of synthetic speech, and inappropriate phrasing can actually make it more difficult under certain conditions for listeners to comprehend synthetic speech.

Paris et al. (2000) also found that prosodic information plays a role in enabling listeners to recall what they have heard in a sentence. These authors investigated the effects of prosody on lexical recall by constructing stimuli both with and without sentence-level prosodic information using both natural and synthetic (DECTalk, SoundBlaster) voices. Paris et al. also looked at the effects of meaningful, sentential contexts on listeners' ability to recall words, and created four different types of stimuli: 1. meaningful sentences with normal intonation; 2. meaningful sentences in monotone; 3. meaningless sentences with normal intonation; 4. meaningless strings of words with no sentence-level prosody. The first two sets of sentences were combinations of Harvard sentences, averaging 15 to 20 words in length; the third set consisted of similar sentences, except with the content words changed to make meaningless sentences (e.g., "Add house before you find the truck..."); and the fourth set consisted of strings of unrelated words (e.g., "In with she plate storm after of proof..."). Listeners were asked to listen to each of these stimuli once and then immediately recall as much of the stimulus as they could. The listeners recalled items from meaningful sentences more easily than items from meaningless strings, and naturally produced words were easier to recall than synthetically produced words. No significant differences were found in recall between DECTalk- and SoundBlaster-produced items. There were several interesting interactions between voice type and presentation context. Table 8 below shows the percentage of lexical items correctly recalled in each condition:

	Condition 1 Normal	Condition 2 No prosody	Condition 3 No meaning	Condition 4 Unstructured
Natural	74	60	51	24
DECTalk	60	60	35	20
SoundBlaster	58	58	34	16

Table 8. Percentage of items correctly recalled, by voice type and presentation context (adapted from Paris et al., 2000).

In the two conditions using normal, sentence-level prosody—Conditions 1 and 3—natural speech items displayed a significant recall advantage over synthetic speech items. This advantage disappeared, however, in Condition 2, in which the listeners heard meaningful sentences with only lexical-level prosody. In Condition 4, there was only a significant difference in correct recall percentages between the natural items and the SoundBlaster items; there was not a significant difference between DECTalk and natural speech. These results indicated, once again, that broad, sentence-level prosodic cues can help listeners not only comprehend speech better, but also help them recall later what they have heard. The results also suggest that the comparative inability of listeners to recall synthetic speech items may not necessarily be due to the impoverished acoustic-phonetic structure of those items, but rather to the failure of the speech synthesis algorithms to incorporate those sentence-level prosodic cues in a natural way.

After the immediate recall task, Paris et al. (2000) played their listeners more samples of sentences produced by the three individual voices and asked them to make subjective judgments of how intelligible and natural-sounding each sentence sample was, on a scale from 1 to 10. The results of this second task are given in Tables 9 and 10.

	Condition 1 Normal	Condition 2 No prosody	Condition 3 No meaning	Condition 4 Unstructured
Natural	9.86	7.80	9.27	5.90
DECTalk	8.20	7.74	6.13	6.07
SoundBlaster	7.10	6.39	5.22	4.11

Table 9. Intelligibility ratings by voice type and presentation context (adapted from Paris et al., 2000).

	Condition 1 Normal	Condition 2 No prosody	Condition 3 No meaning	Condition 4 Unstructured
Natural	9.71	5.58	9.49	5.23
DECTalk	5.78	4.19	4.70	4.27
SoundBlaster	4.60	3.86	3.67	3.05

Table 10. Naturalness ratings by voice type and presentation context (adapted from Paris et al., 2000).

Sentence-level prosody had the same effect on intelligibility ratings as it did on free recall scores—with sentence-level prosody, listeners perceived natural speech tokens as much more intelligible than both types of synthetic speech tokens, but, without sentence-level prosody, there was no significant perceived intelligibility difference between the natural and DECTalk voices (and SoundBlaster was still rated as less intelligible than the other two voices). Sentence-level intonation played an even more significant role in the naturalness judgments, where the natural stimuli were rated considerably higher than the synthetic stimuli in Conditions 1 and 3, and still slightly higher than the synthetic stimuli in Conditions 1 and 2.

These results demonstrate that the perceived intelligibility of sentence stimuli corresponds closely to the listeners' ability to recall the content words from these sentences. Since sentence-level prosody influenced both intelligibility and naturalness ratings, it is likely that it is necessary to incorporate appropriate prosodic patterns into synthetic speech for it to match the intelligibility and retention of natural speech in all listening conditions. Paris et al.'s (2000) subjective naturalness ratings also indicated that sentence-level prosody dictates—more than any other factor—the potential level of perceived naturalness in synthetic speech stimuli. Since Paris et al.'s natural tokens were still judged to be more natural than synthetic tokens in conditions which lacked sentence-level prosody, however, some sub-prosodic features of human speech evidently contribute to perceived naturalness as well, such as glottal source characteristics, as noted by Nusbaum et al. (1995).

Summary. The available research has shown that the naturalness of synthetic speech depends on a variety of factors, including segmental intelligibility, the pragmatic appropriateness of particular pitch contours, and the amount of variability in the implementation of synthetic prosody. Assessing the naturalness of synthetic speech independently of these factors has proven difficult, and attempts to do so have primarily focused on evaluating the naturalness of isolated voice source information in the synthetic speech signal. Research has also shown that appropriate prosodic information may facilitate the comprehension and recall of sentences produced synthetically. Together, the findings of this body of research suggest that generating more naturalistic and appropriate sentence-level prosody will be an

important research challenge in the effort to develop more natural-sounding, highly intelligible synthetic speech in the years to come.

Conclusions and Directions for Future Research

Research on the perception of synthetic speech has always had both practical and theoretical motivations. Practically, research on the perception of synthetic speech can reveal what limitations there are on the comprehension of synthetic speech in applied settings, and what work needs to be done in order to overcome those limitations and improve the quality of speech synthesis. Theoretically, research on the perception of synthetic speech is important because it offers a unique window into how human listeners can deal with and extract information from unnatural and impoverished speech signals. By comparing the perception of such impoverished speech signals with natural speech, researchers can obtain new fundamental knowledge about which aspects of natural speech are important to the robust perceptual abilities of human listeners in a wide variety of listening environments.

The practical implications of research findings on the perception of synthetic speech over the past 30+ years are clear. Research on the perception of synthetic speech in noise, for instance, has revealed that, even though the segmental intelligibility of current speech synthesis systems may approximate natural speech under clear listening conditions, speech perception deteriorates rapidly and significantly in noisy listening conditions. This finding suggests that synthetic speech applications would be of limited utility in e.g., noisy work environments, in football stadiums, or even over a cell phone. Investigating ways to overcome the persistent limitations on synthetic speech intelligibility, however, provides several promising opportunities for future research. For instance, audio-visual speech synthesis systems provide a possible solution to the practical problem of improving the intelligibility of synthetic speech in noise (Cohen & Massaro, 1994; Ezzat & Poggio, 2000). The information in the speech signal that people can perceive in the visual domain is largely complementary to that which they can perceive in the auditory domain (Calvert, Spence & Stein, 2004; Massaro, 1997; Summerfield, 1987) and furthermore, provides robust cues to those features of speech which are most often misperceived in noisy listening conditions. Visual information about speech has been shown to provide substantial gains in intelligibility when audio-visual stimuli are presented in noise (Sumby & Pollack, 1954). Although viable systems for producing synthetic audio-visual speech have been around for some time, little is known about people's ability to perceive the visual information in the synthetic speech tokens that these systems produce. Is the visual-only perception of synthetic speech significantly different from the visual-only perception of natural speech? And, does synthetic visual speech boost the intelligibility of synthetic speech in noise as much as natural visual cues help improve the intelligibility of natural speech in noise? Answering basic research questions such as these may prove important not only to our future understanding of the utility of audio-visual speech synthesis, but also to our understanding of the fundamental characteristics that are important for multimodal perception of natural speech.

Research on the perception of synthetic speech also indicates that it requires more cognitive resources than the perception of natural speech. This finding suggests that synthetic speech applications might be of limited utility to listeners who are engaged in cognitively demanding tasks such as driving a car, flying an airplane, directing air traffic, etc. One way to limit such demands on listeners is to reduce the number of possible messages that a synthetic speech system can transmit. In diagnostic tests, such as the MRT, listeners can correctly identify synthetic speech tokens nearly as well as natural speech tokens when they only have to select responses from a small, limited set of response alternatives. In more cognitively demanding environments, therefore, synthetic speech may be better suited to transmitting only a small number of messages to listeners—such as, for example, warnings in an airplane cockpit (Simpson & Williams, 1980), or digits or letters of the alphabet. However, the production of a limited set of messages could be more easily handled by a recorded database of those messages, rather than a text-to-

speech system which is designed to produce an unlimited number of messages (see Allen, Hunnicutt & Klatt, 1987).

The potential for success of any synthetic speech application also depends on the listener who is using the system. Research on the perception of synthetic speech has shown that alternative groups of listeners—e.g., children, non-native speakers, and listeners with hearing impairment—have more difficulty understanding synthetic speech than the average, normal-hearing, college-aged native speaker does. The ability of all groups of listeners, however, to extract information from synthetic speech tends to improve the more they are exposed to it. The limits of such improvement on the perception of synthetic speech are not precisely known. Improvement in the ability to perceive synthetic speech nonetheless has, by its very nature, the potential to create another alternative group of listeners. “Expert listeners,” for instance, may emerge among people who listen to particular forms of synthetic speech for long periods of time on a daily basis. Such “expert listeners” may include, for instance, blind people who regularly use applications on their computers which output textual information to them in the form of synthetic speech. Such listeners would likely have far more experience listening to synthetic speech than any one listener might receive in a perceptual training experiment in the laboratory, and they may have thus developed correspondingly better abilities to extract acoustic-phonetic information from synthetic speech signals. Research on this potential group of listeners may reveal what upper limits (if any) exist on the fundamental human ability to extract information from synthetic speech, and may also determine if extensive amounts of listening experience can close the gap in intelligibility between synthetic and natural speech. It may also be instructive to investigate the extent to which expert knowledge of one form of synthetic speech may improve the perception of other forms of synthetic speech.

The finding that synthetic speech has been consistently shown to be less intelligible and less comprehensible than natural speech—across a wide variety of testing conditions and listener populations—has led researchers to draw a number of important theoretical conclusions about the aspects of natural speech which facilitate robust speech perception. These conclusions emerged from a consideration of how natural speech differs from synthetic speech that was produced by rule, since this was the form of synthetic speech that has been most commonly tested in synthetic speech perception research. Even very high quality synthetic speech produced by rule lacks both the rich variability and acoustic-phonetic cue redundancy characteristic of natural speech, and it is presumably the absence of these dynamic natural speech characteristics—along with the lack of appropriate prosodic information—which makes synthetic speech produced by rule difficult for listeners to perceive and comprehend.

However, not all synthetic speech is produced strictly by rule. “Concatenative” speech synthesis techniques, for instance, use natural human utterances as a voice source. Concatenative synthesis thus has a natural-sounding quality which has helped increase its popularity and use in recent years. Despite this increase in popularity, however, relatively little is known about the perception of speech produced by concatenative synthesis techniques and how it may differ from either synthetic speech produced by rule or natural speech. Using natural speech utterances as source material should improve the quality of synthetic speech in some putatively important respects—such as incorporating robust and redundant sets of cues to individual segments into the signal, for instance. However, using natural speech utterances as source material may also potentially damage the quality of synthetic speech in other ways—by introducing, for example, perceptible discontinuities between two adjoining units in the concatenated speech signal. It is perhaps not surprising, therefore, that existing research on the intelligibility of concatenative versus formant synthesis (e.g., Venkatagiri, 2003) indicates that concatenative synthesis produces highly intelligible consonants—for which it preserves the natural cues and formant transitions—but somewhat less intelligible vocalic segments, where the discontinuities between the concatenated source units typically exist.

The recent findings reported by Venkatagiri (2003) confirm that the acoustic-phonetic cue redundancy in natural speech is, indeed, important to perception. More research needs to be done, however, not only on the differences which may exist in perception between concatenative and formant synthesis—especially with respect to the vast body of research findings from the past 30+ years—but also on the specific role that acoustic-phonetic variability plays in speech perception. “Variability” has a wide variety of sources in speech—e.g., speaker, dialect, gender, age, emotional state, social identity, etc.—and little is known about what role (if any) these different sources of variability may play in facilitating the perception of natural speech. Incorporating different sources of variability into synthetic speech is a logical way to test their effects on speech intelligibility, and it may also provide a potential means of improving the overall intelligibility of synthetic speech. (Stevens, 1996)

Although the intelligibility of high quality speech synthesis systems currently approximates that of natural speech in clear listening conditions, improving its intelligibility in adverse listening conditions still remains an important research goal. One increasingly popular application of speech synthesis is providing spoken output of driving directions from computerized, on-board navigation systems in automobiles. The drivers who need to understand these spoken directions must do so in frequently noisy listening conditions and under a significant cognitive load. Research has shown that the perception of even high quality synthetic speech deteriorates significantly under such listening conditions. Determining how to improve the intelligibility of synthetic speech in noisy and cognitively demanding listening conditions should therefore help improve the viability and safety of such applications of speech synthesis.

Improving the quality of synthetic speech may also require a better understanding of what makes synthetic speech sound “unnatural” to most human listeners. However, even after three decades of research on the perception of synthetic speech, little is known about “naturalness” and its relationship to speech perception. Most research on the perception of synthetic speech has focused, instead, on studying the segmental intelligibility and comprehension of synthetic speech in comparison to natural speech. Now that the segmental intelligibility of synthetic speech has reached near-natural levels, however, determining what makes one kind of synthetic speech sound more natural or preferable to listeners should become an increasingly important research goal.

It is likely that incorporating appropriate prosodic contours into synthetic speech will increase its perceived “naturalness.” Paris et al. (2000) found, for instance, that removing sentence-level prosody from natural speech not only diminished its perceived naturalness but also reduced its perceived intelligibility to a level comparable to that of synthetic speech. Listeners’ ability to interpret synthetic speech therefore seems to depend to some extent on the presence of appropriate prosodic cues in synthetic speech. Testing the perceived “appropriateness” of particular prosodic contours, however, may prove more difficult because listeners cannot, in general, describe the prosodic information they perceive in linguistic terms. Thus, it may not be possible to investigate the perception of prosody in synthetic speech directly; instead, future research in this area may have to investigate the perception of prosody using indirect methods by focusing on the effects that prosodic structure has on, e.g., measures of memory, attention, processing load and processing speed, in quiet and noise, under a wide range of listening conditions, both with and without cognitive loads. Such research may require the development of entirely new assessment methods in order to better understand how prosodic information can facilitate or inhibit these elements of the perception of synthetic speech (Pisoni, 1997).

Developing new experimental methods which can target higher-level comprehension processes may also enable researchers to investigate and identify the scope of the detrimental effects of synthetic speech—as compared to natural speech—on the human language comprehension system. Duffy and Pisoni (1992) logically suggested that the phonological encoding of speech stimuli must occur before the interpretation of message-level information in comprehension, even though processing at both levels

could go on concurrently. Synthetic speech appears to make processing more difficult at both the phonological and message levels; however, it is unclear whether this is due solely to the difficulties inherent in interpreting synthetic speech at the phonetic level—which may have cascading effects on subsequent stages in the comprehension process—or due to independent difficulties in the processing of synthetic speech that emerge at the stage of message-level semantic interpretation. Developing new behavioral tasks and better assessment methods which can target higher-level comprehension processes, independently of low-level phonemic encoding effects, could help clarify whether the difficulty in perceiving synthetic speech exists at the level of segmental interpretation alone, or whether it also causes specific, independent problems for semantic processing. Examining how these distinct levels of processing interact with one another in the perception of synthetic speech may shed light on how these levels operate together in the comprehension of natural speech as well.

Considerations such as these on the future directions of research on the perception of synthetic speech reflect the fact that speech scientists will likely have to develop more sophisticated methods as speech synthesis technology continues to improve. No matter what technological developments may come to pass, however, research on synthetic speech perception will continue to have direct practical benefits for speech synthesis technology, as well as provide a unique opportunity for speech scientists to investigate what makes synthetic speech hard for human listeners to understand and which aspects of natural speech help them perform the task of normal speech perception so well.

References

- Allen, J., Hunnicutt, M.S. & Klatt, D. (1987). *From Text to Speech: The MITalk System*. New York: Cambridge University Press.
- Atal, B.S. & Hanauer, S.L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50, 637-655.
- Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.
- Calvert, G., Spence, C. & Stein, B.E. (2004). *The Handbook of Multisensory Processing*. Cambridge, MA: MIT Press.
- Carlson, R., Granstrom, B. & Larsson, K. (1976). Evaluation of a text-to-speech system as a reading machine for the blind. *Quarterly Progress and Status Report, STL-QPSR 2-3*. Stockholm: Royal Institute of Technology, Department of Speech Communications.
- Clark, J.E. (1983). Intelligibility comparisons for two synthetic and one natural speech source. *Journal of Phonetics*, 11, 37-49.
- Cohen, M.M. & Massaro, D.W. (1994). Synthesis of visible speech. *Behavior Research Methods, Instruments and Computers*, 22, 260-263.
- Duffy, S.A. & Pisoni, D.B. (1991). Effects of sentence context on the signal duration required to identify natural and synthetic words. In *Research on Speech Perception Progress Report No. 17*, (Pp. 341-354). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Duffy, S.A. & Pisoni, D.B. (1992). Comprehension of synthetic speech produced by rule: a review and theoretical interpretation. *Language and Speech*, 35(4), 351-389.
- Dutoit, T. & Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13, 435-440.
- Egan, J.P. (1948). Articulation testing methods. *Laryngoscope*, 58, 955-991.
- Ezzat, E. & Poggio, T. (2000). Visual Speech Synthesis by Morphing Visemes. *International Journal of Computer Vision*, 38, 45-57.
- Fairbanks, G. (1958). Test of phonemic differentiation: the rhyme test. *Journal of the Acoustical Society of America*, 30, 596-600.
- Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

- Greene, B.G. (1983). Perception of synthetic speech by children. In *Research on Speech Perception Progress Report No. 9*, (Pp. 335-348). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Greene, B.G. (1986). Perception of synthetic speech by nonnative speakers of English. In *Proceedings of the Human Factors Society*, 1340-1343. Santa Monica, CA.
- Greene, B.G., Manous, L.M. & Pisoni, D.B. (1984). Perceptual evaluation of DECTalk: a final report on version 1.8. In *Research on Speech Perception Progress Report No. 10*, (Pp. 77-128). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Greene, B.G. & Pisoni, D.B. (1988). Perception of synthetic speech by adults and children: research on processing voice output from text-to-speech systems. In L.E. Bernstein (Ed.), *The Vocally Impaired: Clinical Practice and Research*, (pp. 206-248). Philadelphia: Grune & Stratton.
- Greenspan, S.L., Nusbaum, H.C. & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14, 421-433.
- Groner, G.F., Bernstein, J., Ingber, E., Pearlman, J. & Toal, T. (1982). A real-time text-to-speech converter. *Speech Technology*, 1, 73-76.
- Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28, 267-283.
- Hoover, J., Reichle, J., Van Tasell, D. & Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*, 30, 425-431.
- House, A.S., Williams, C.E., Hecker, M.H.L. & Kryter, K.D. (1965). Articulation-testing methods: consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37, 158-166.
- Humes, L.E., Nelson, K.J. & Pisoni, D.B. (1991). Recognition of synthetic speech by hearing-impaired elderly listeners. *Journal of Speech and Hearing Research*, 34, 1180-1184.
- Humes, L.E., Nelson, K.J., Pisoni, D.B. & Lively, S.E. (1993). Effects of age on serial recall of natural and synthetic speech. *Journal of Speech and Hearing Research*, 34, 1180-1184.
- Hustad K.C., Kent R.D. & Beukelman D.R. (1998). DECTalk and MacinTalk speech synthesizers: intelligibility differences for three listener groups. *Journal of Speech Language and Hearing Research* 41, 744-752.
- Ingemann, F. (1978). Speech synthesis by rule using the FOVE program. In *Haskins Laboratories Status Report on Speech Research, SR-54*, (pp. 165-173). New Haven, CT: Haskins Laboratories.
- Kalikow, D.N., Stevens, K.N. & Elliott, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61(5), 1337-1351.
- Kangas, K.A. & Allen, G.D. (1990). Intelligibility of synthetic speech for normal-hearing and hearing-impaired listeners. *Journal of Speech and Hearing Disorders*, 55, 751-755.
- Koul, R., & Allen, G. (1993). Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in presence of noise. *Journal of Speech and Hearing Research*, 36, 790-798.
- Logan, J.S., Greene, B.G. & Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86, 566-581.
- Luce, P.A. (1981). Comprehension of fluent synthetic speech produced by rule. In *Research on Speech Perception Progress Report No. 7*, (Pp. 229-242). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Luce, P.A., Feustel, T.C. & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25, 17-32.
- Luce, P.A. & Pisoni, D.B. (1983). Capacity-demanding encoding of synthetic speech in serial-ordered recall. In *Research on Speech Perception Progress Report No. 9* (pp. 295-309). Bloomington, IN: Speech Research Laboratory, Indiana University.

- Manous, L.M. & Pisoni, D.B. (1984). Effects of signal duration on the perception of natural and synthetic speech. In *Research on Speech Perception Progress Report No. 10* (pp. 311-321). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Manous, L.M., Pisoni, D.B., Dedina, M.J., & Nusbaum, H.C. (1985). Comprehension of natural and synthetic speech using a sentence verification task. In *Research on Speech Perception Progress Report No. 11* (pp. 33-57). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Massaro, D.W. (1997). *Perceiving Talking Faces: from Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
- Mattingly, I.G. (1968). *Synthesis by rule of General American English*. Ph.D. dissertation, Yale University. (Issued as supplement to Haskins Laboratories Status Report on Speech Research.)
- Miranda, P. & Beukelman, D.R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication, 3*, 120-128.
- Miranda, P. & Beukelman, D. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication, 6*, 61-68.
- Mitchell, P. & Atkins, C. (1988). A comparison of the single word intelligibility of two voice output communication aids. *Augmentative and Alternative Communication, 4*, 84-88.
- Moody, T. & Joost, M. (1986). Synthesized speech, digitized speech, and recorded speech: a comparison of listener comprehension rates. In *Proceedings of the Voice Input/Output Society*. Alexandria, VA.
- Moulines, E. & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453-467.
- Nusbaum, H.C., Dedina, J.J. & Pisoni, D.B. (1984). Perceptual confusions of consonants in natural and synthetic CV syllables. In *Speech Research Laboratory Technical Note, 84-02*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 1*, 7-19.
- Nusbaum, H.C. & Pisoni, D.B. (1984). Perceptual evaluation of synthetic speech generated by rule. In *Proceedings of the 4th Voice Data Entry Systems Applications Conference*. Sunnyvale, CA: Lockheed.
- Nusbaum, H.C., Schwab, E.C. & Pisoni, D.B. (1984). Subjective evaluation of synthetic speech: measuring preference, naturalness and acceptability. In *Research on Speech Perception Progress Report No. 10* (pp. 391-408). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Nye, P.W. & Gaitenby, J.H. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). In *Haskins Laboratories Status Report on Speech Research, SR-33*, 77-91. New Haven, CT: Haskins Laboratories.
- Nye, P.W. & Gaitenby, J.H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. In *Haskins Laboratories Status Report on Speech Research, SR-37/38*, 169-190. New Haven, CT: Haskins Laboratories.
- Nye, P.W., Ingemann, F. & Donald, L. (1975). Synthetic speech comprehension: a comparison of listener performances with and preferences among different speech forms. In *Haskins Laboratories Status Report on Speech Research, SR-41*, 117-126. New Haven, CT: Haskins Laboratories.
- Paris, C.R., Gilson, R.D., Thomas, M.H. & Silver, N.C. (1995). Effect of synthetic voice intelligibility upon speech comprehension. *Human Factors, 37*, 335-340.
- Paris C.R., Thomas M.H., Gilson R.D., & Kincaid J.P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors 42*, 421-431.
- Pisoni, D.B. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America, 70*, S98.

- Pisoni, D.B. (1987). Some measures of intelligibility and comprehension. In J. Allen, M.S. Hunnicutt, & D.H. Klatt (Eds.), *From Text to Speech: the MITalk System*, Pp 151-171. Cambridge, UK: Cambridge University Press.
- Pisoni, D.B. (1997). Perception of synthetic speech. In J.P.H. van Santen, R.W. Sproat, J.P. Olive & J. Hirschberg (Eds.), *Progress in Speech Synthesis*, Pp 541-560. New York: Springer-Verlag.
- Pisoni, D.B. (1997). Some Thoughts on "Normalization" in Speech Perception. In K. Johnson & J. W. Mullennix (eds.), *Talker Variability in Speech Processing*, (pp. 9-32). San Diego: Academic Press.
- Pisoni, D.B. & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In *1980 IEEE International Conference on Acoustics, Speech and Signal Processing*, 572-575. New York: IEEE.
- Pisoni, D.B. & Koen, E. (1981). Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. In *Research on Speech Perception Progress Report No. 7* (pp. 243-254). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Pisoni, D.B., Manous, L.M., & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, 2, 303-320.
- Pisoni, D.B., Nusbaum, H.C., & Greene, B.G. (1985). Perception of synthetic speech generated by rule. In *Proceedings of the Institute of Electrical and Electronics Engineers*, 73 (11), 1665-1675.
- Pols, L.C.W. (1989). Assessment of text-to-speech synthesis systems. In A.J. Fourcin, G. Harland, W. Barry & V. Hazan (eds.), *Speech Input and Output Assessment*, (pp. 55-81). Chichester, England: Ellis Horwood.
- Pols, L.C. W. (1992). Quality assessment of text-to-speech synthesis by rule. In S. Furui & M.M. Sondhi (eds.), *Advances in Speech Signal Processing*, (pp. 387-416). New York: Marcel Dekker.
- Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G. & Mullennix, J.W. (1991). Comprehension of synthetic speech produced by rule: word monitoring and sentence-by-sentence listening times. *Human Factors*, 33, 471-491.
- Reynolds, M.E., Bond, Z.S. & Fucci, D. (1996). Synthetic speech intelligibility: comparison of native and non-native speakers of English. *Augmentative and Alternative Communication*, 12, 32-36.
- Reynolds, M.E. & Fucci, D. (1998). Synthetic speech comprehension: a comparison of children with normal and impaired language skills. *Journal of Speech, Language and Hearing Research*, 41, 458-466.
- Reynolds, M.E., Isaacs-Duvall C. & Haddox M.L. (2002). A comparison of learning curves in natural and synthesized speech comprehension. *Journal of Speech Language and Hearing Research* 45, 802-820.
- Reynolds, M.E., Isaacs-Duvall, C., Sheward, B. & Rotter, M. (2000). Examination of the effects of listening practice on synthesized speech comprehension. *Augmentative and Alternative Communication*, 16, 250-259.
- Reynolds, M.E. & Jefferson, L. (1999). Natural and synthetic speech comprehension: comparison of children from two age groups. *Augmentative and Alternative Communication*, 15, 174-182.
- Rounsefell, S., Zucker, S.H. & Roberts, T.G. (1993). Effects of listener training on intelligibility of augmentative and alternative speech in the secondary classroom. *Education and Training in Mental Retardation*, 28, 296-308.
- Rupprecht, S., Beukelman, D. & Vrtiska, H. (1995). Comparative Intelligibility of five synthesized voices. *Augmentative and Alternative Communication*, 11, 244-247.
- Sanderman, A.A. & Collier, R. (1996). Prosodic rules for the implementation of phrase boundaries in synthetic speech. *Journal of the Acoustical Society of America* 100, 3390-3397.
- Sanderman, A.A. & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, 40, 391-409.

- Schwab, E.C., Nusbaum, H.C. & Pisoni, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.
- Simpson, C.A. & Williams, D.H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. *Human Factors*, 22, 319-330.
- Slowiaczek, L.M. & Nusbaum, H.C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, 27, 701-712.
- Slowiaczek, L.M. & Pisoni, D.B. (1982). Effects of practice on speeded classification of natural and synthetic speech. In *Research on Speech Perception Progress Report No. 7*, (pp. 255-262). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Stevens, K.N. (1996). Understanding variability in speech: a requisite for advances in speech synthesis and recognition. *Journal of the Acoustical Society of America*, 100, 2634.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (eds.), *Hearing by Eye: The Psychology of Lipreading*, (pp. 3-51). Hillsdale, NJ: Lawrence Erlbaum & Associates.
- Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- Sutton, B., King, J., Hux, K., & Beukelman, D. (1995). Younger and older adults' rate performance when listening to synthetic speech. *Augmentative and Alternative Communication*, 11, 147-153.
- Terken, J. (1993). Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language*, 7, 27-48.
- Terken, J. & Lemeer, G. (1988). Effects of segmental quality and intonation on quality judgments for texts and utterances. *Journal of Phonetics*, 16, 453-457.
- van Santen, J.P.H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8, 95-128.
- Venkatagiri, H.S. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication*, 10, 96-104.
- Venkatagiri, H.S. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America* 113, 2095-2104.