INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 3, 15–25 (1999)

The Effect of Lexical Complexity on Intelligibility

ALEXANDER L. FRANCIS AND HOWARD C. NUSBAUM
Department of Psychology, University of Chicago, 5848 S. University Ave., Chicago, IL 60637
alfr@speech.uchicago.edu
hcn@speech.uchicago.edu

Received June 17, 1997; Accepted May 8, 1999

Abstract. Most intelligibility tests are based on the use of monosyllabic test stimuli. This constraint eliminates the ability to measure the effects of lexical stress patterns, complex phonotactic organizations, and morphological complexity on intelligibility. Since these aspects of lexical structure affect speech production (e.g., by changing syllable duration), it is likely that they affect the structure of acoustic-phonetic patterns. Thus, to the extent that text-to-speech systems fail to modify acoustic-phonetic patterns appropriately in polysyllabic words, intelligibility may suffer. This means that while most standard intelligibility tests may accurately estimate the intelligibility of monosyllabic words, this estimate may not generalize as well to predict the intelligibility of words with more complex lexical structures. The present study was carried out to measure how words varying in lexical complexity differ in intelligibility. Monosyllabic, bisyllabic, and polysyllabic words were used varying in morphological complexity (monomorphemic or polymorphemic). Listeners transcribed these stimuli spoken by two human talkers and two text-to-speech systems varying in speech quality. The results indicate that lexical complexity does affect the measured intelligibility of synthetic speech and should be manipulated in order to accurately predict the performance of text-to-speech systems with unrestricted natural text.

Keywords: text-to-speech, intelligibility assessment, lexical complexity

Introduction

Under most circumstances, it is more difficult to recognize and understand synthetic speech than natural speech (Nusbaum and Pisoni, 1985; Ralston et al., 1995). First, recognition of spoken words produced by a text-to-speech system is much less accurate than recognition of words produced by a human talker (cf. e.g., Logan et al., 1989). Furthermore, these accuracy differences between synthetic speech and natural speech are found at the level of phoneme perception (Spiegel et al., 1990), word recognition in isolated words (Logan et al., 1989) and in sentences (Slowiaczek and Nusbaum, 1985) and discourse comprehension (Ralston et al., 1995). Second, recognition of synthetic speech requires more effort and attention than recognition of natural speech (Luce et al., 1983). Listeners have to work harder to recognize synthetic speech. This means that synthetic speech may interfere more with the performance of other tasks (e.g., flying a plane or remembering information from a database) than would natural speech.

Certainly synthetic speech does not provide all the acoustic-phonetic cues that listeners generally expect based on their experience with natural speech. The phonetic implementation rules of text-to-speech systems use only those acoustic cues that have been identified from acoustic analyses of natural speech (e.g., see Nusbaum and Pisoni, 1985). Moreover, the covariation among cues in those rules is at best only a poor approximation of natural speech. Even worse, there may be actual errors in the rules such that in some phonetic contexts the acoustic cues are inappropriate for the intended phonetic segment, which can mislead the

listener. Thus, the quality of acoustic-phonetic patterns in synthetic speech is limited by the rules that govern the process of converting text to speech. Since these rules may not provide all the information deployed by human talkers in conveying a message and sometimes even may provide incorrect information, they constitute a significant limitation on the acoustic-phonetic pattern structure of synthetic speech.

The measured accuracy of recognizing speech produced by a text-to-speech system reflects both the way the system encodes linguistic information into acoustic patterns and the knowledge and expectations the listener brings to bear on recognizing those patterns. Thus, intelligibility tests measure how well a text-to-speech system *models* the way human talkers encode a wide range of linguistic structures into the acoustic patterns of speech, since this is the basis for the rules that govern the synthesis process as well as the basis for the listeners' perceptual expectations. Therefore, intelligibility tests should be sensitive to all the factors that affect the process of speech production. This means that tests should measure the effects of different local phonetic environments, stress levels and patterns, speaking rate differences, etc. If several aspects of variation in speech production are used in an application (e.g., fast vs. slow speech, a large vocabulary, different syntactic structures that result in different intonation patterns) and these aspects affect the relationship between linguistic units and the acoustic patterns of speech, it is important to measure the effects of this variation in an intelligibility test. Unfortunately, there are many such factors that have not been systematically investigated to determine whether they do affect intelligibility.

One important example is the effect of lexical structure on intelligibility. From the perspective of production, word-level prosody related to morphological structure causes significant restructuring of the acoustic signal of words. For example, although adding suffixes to a root morpheme generally decreases the absolute duration of the root, the suffix *-iness*, as in *speediness*, causes a greater decrease in root duration than does the similarly bisyllabic affix *-ily*, as in *speedily*. Lehiste (1972) proposed that this is due to the derived nature of the suffix *-iness*, in that *speediness* can be thought of as being derived from *speedy* + *-ness*, where *speedy* is already derived by adding *-y* onto *speed*. It is possible that listeners could use their knowledge of the effects of derivational morphemes on acoustic parameters such as syllable length in order to facilitate word recognition. Indeed, Grosjean and Gee (1987) showed that similar kinds of prosodic information may play a role during word recognition. Thus, if the acoustic effects of morpheme combination are not encoded into speech appropriately, listeners may make errors in recognition.

The morphological structure of words may also play a role in how easy they are to recognize. Even when listening to natural speech, when listeners are confronted with a noisy signal they are able to make use of their morphological knowledge to narrow the search space of possible interpretations of a poorly heard word. In such situations, morphologically complex words may be easier to recognize, because their greater complexity provides more structural cues to their identity, even when some segmental cues are obscured. However, if the acoustic-phonetic cue patterns of a speech signal are sufficiently uninformative or misleading, as in the case of a poor text-to-speech system, listeners may not be able to obtain enough segmental information about a word to develop an accurate mental representation of the word's morphological structure. Thus, the benefits of listening to morphologically complex words may only be manifest in natural speech and synthetic speech of sufficiently high quality. Unfortunately, most current intelligibility tests focus on measuring intelligibility in a monosyllabic context (see Schmidt-Nielsen, 1995; Spiegel et al., 1990). Therefore these tests overlook the possible effects of lexical complexity on recognition, both in terms of production and perception.

Another factor that is likely to affect intelligibility that is not reflected in standard monosyllabic test corpora is word length. Increasing the number of syllables in a word improves the robustness of word and phoneme recognition (see Cole and Rudnicky, 1983; Grosjean, 1980; Samuel, 1981). Thus, tests of intelligibility using only monosyllabic words may underestimate the overall intelligibility of a given speech synthesizer. On the other hand, the prosodic demands of increasing word length may require a significant restructuring of the pronunciation of particular segments, especially vowels, in a given word, as discussed by Lehiste (1972). Furthermore, it has been proposed that prosodic characteristics such as alternating stress may play a significant role in word recognition (Grosjean and Gee, 1987). If a speech synthesizer does not accurately reproduce these word-length related aspects of human speech, intelligibility may suffer. However, monosyllabic words cannot vary in prosodic characteristics such as stress placement, and therefore they are of little use in testing the contributions of these factors to intelligibility.

Even if the rules of a text-to-speech system do attempt to incorporate the effects of morphological complexity and word length on acoustic patterns, the acoustic-phonetic patterns of synthetic speech still may not match (and may in fact mislead) the expectations of listeners because the acoustic consequences of these factors are still only incompletely understood (Nusbaum and Pisoni, 1985). As a result, lexically complex materials produced by a text-to-speech system may be more poorly specified acoustically than monomorphemic, monosyllabic words. Thus, the difference in recognition accuracy between words spoken by a human and those produced by a text-to-speech system may be greater for lexically complex words than for simpler tokens. If this is the case, estimates of segmental intelligibility based on monosyllabic materials might overestimate the performance of particular speech synthesizers.

In addition, monosyllabic, monomorphemic words offer little opportunity for listeners to use lexical knowledge to aid in recognition. It is well known that meaningful contexts provide a great deal of aid in recognizing spoken words. For example, words spoken in syntactically correct, semantically unambiguous sentences are easier to understand than those presented in syntactically correct, semantically anomalous sentences (Nusbaum and Pisoni, 1985). Similarly, words with complex morphology may provide listeners with more contextual cues for phoneme identification than do simpler, monomorphemic words. If this is the case, estimates of intelligibility based on monosyllabic materials might underestimate the intelligibility of multisyllabic, multimorphemic words produced by particular speech synthesizers. Finally, monomorphemic, monosyllabic test corpora cannot reveal any differences that might exist in the intelligibility of morphologically complex speech produced by two different text-to-speech systems.

The present study was carried out to investigate how lexical structure affects segmental intelligibility. We compared word recognition performance for natural and synthetic speech varying in word length and morphological complexity. Since increasing the number of syllables in a word improves the robustness of word and phoneme recognition, increasing the number of syllables should improve recognition of synthetic speech more than recognition of natural speech. However, if synthetic speech is impoverished in terms of the manner or degree in which morphological complexity affects the acoustic-phonetic structure of an utterance, increasing morphological complexity could reduce segmental intelligibility for synthetic speech compared to natural speech.

Method

Subjects

Thirteen subjects participated in this experiment. All subjects were native speakers of a North American dialect of English, and none had any history of speaking or hearing disability. Subjects were recruited from the student population of the University of Chicago and were paid $8 each for approximately 45 minutes of participation.

Stimuli

Figure 1 shows the basic design of the stimuli used in this experiment. Stimuli varied in number of syllables so that there were one, two, and three or four syllable words. These will be referred to as the *monosyllabic*, *bisyllabic*, and *polysyllabic* conditions. Stimuli also varied in morphological complexity, falling into either the *monomorphemic* or *polymorphemic* categories (though only bisyllabic and polysyllabic words could be polymorphemic; all monosyllabic words were also monomorphemic).

As is shown in Fig. 1, two-syllable words could consist of one or two constituent morphemes. It is important to note that, though some bisyllabic and polysyllabic words such as *donut* (*doughnut*) contain syllables which constitute distinct morphemes in isolation, in the context of the word in question the syllables do not function as separate morphemes, and therefore they are not expected to affect the suprasegmental pattern of the word in the same manner as if they were distinct morphemes (e.g., in *bread dough*). Similarly, although some words such as *vesicle* contain sound sequences that are morpheme-like in that they appear

	Monosyllabic	**Bisyllabic**	**Polysyllabic**
Monomorphemic	dole vest	donut versus	dolomite vesicle
Polymorphemic		doleful vestment	dolefully vestmented

Figure 1. Examples of members of the five categories of words used in the experiment.

frequently in English (such as *-icle*), they do not function in these words as morphological affixes. Finally, although many of the words we classify as monomorphemic are etymologically more complex (e.g., *proclaim*), this historical complexity probably does not affect online perceptual processing, at least for most native speakers' intuitions.

Polysyllabic words could either be monomorphemic or polymorphemic. We selected polymorphemic words that contained as many morphemes as syllables. Monomorphemic and polymorphemic items were roughly matched according to their overall sound patterns, and, when possible, polymorphemic tokens were constructed using monomorphemic tokens (even when the derivation is not in fact etymologically true). For example, *vest*, *versus*, and *vesicle* were considered matched because they all start with [v], followed by a mid vowel and an [s], and, using *vest*, it was possible to construct *vestment* and *vestmented*. Although these words are not, in themselves, likely to be familiar to listeners or in the pronunciation dictionary of a computer speech synthesizer, they are derived from familiar words using highly productive derivational morphemes, and in spoken English such derivations are quite natural and common. If a speech synthesizer is not able to produce the acoustic consequences of morphological derivations in concordance with listeners' expectations, then it is important to be able to determine whether this inability has an effect on the intelligibility of such derived words. The complete set of words used in this experiment is included in Appendix A.

The stimuli were produced by two male human talkers and two text-to-speech systems (Votrax Type-n-Talk and DECTalk 4.2) which differ in terms of their measured intelligibility on standard monosyllabic test corpora (see Logan et al., 1989). The text-to-speech systems were operated using their default settings for speaking rate and talker characteristics. For this study we chose to use the older synthesizers because in this paper we are more concerned with the methods of evaluation, rather than the actual evaluation itself. Thus, we preferred to use synthesizers for which performance statistics are well known, such that our results are more indicative of the success of the evaluation methods we are proposing, and less a consequence of the specific capabilities of a particular synthesizer. Also for these reasons the responses to the two human talkers were combined, and all statistics were performed using this "averaged" human talker, rather than the individual talkers' raw scores.

The synthetic speech was digitized directly from the output audio jack of the synthesizing computer at a sampling rate of 10 kHz and a quantization rate of 12 bits after low-pass filtering at 4.8 kHz. The human talkers were also recorded in a quiet room directly to digital sound files. They were also digitized at a 10 kHz sampling rate, with a quantization rate of 12 bits after low- pass filtering at 4.8 kHz. All stimuli were digitally reduced to a mean RMS amplitude of 60 dB over the entire token to eliminate amplitude variation between sets of words. Stimuli were presented binaurally via high quality headphones in a sound-treated environment at a comfortable listening level (76 dB SPL).

Procedure

Subjects were tested in groups varying in size from one to three subjects. Each subject identified the spoken words in three blocks of trials. These blocks were constructed to control the number of syllables in each word in the block (monosyllabic, bisyllabic, or polysyllabic). Within each block, subjects identified words produced by all four talkers (two natural and two synthetic) in random order. Within the bisyllabic and polysyllabic blocks, words varied in morphological complexity. Thus some subjects heard the monosyllabic (and therefore necessarily monomorphemic) words in the first block, and bisyllabic (monomorphemic and multimorphemic) words in the second block, and finally polysyllabic (monomorphemic and multimorphemic) words in the third block. The order of blocks was counterbalanced across subjects, but the order of words within each block was randomized.

Subjects were seated at computers in individual sound-attenuating booths. They were told that they would be hearing speech produced by computer speech synthesizers as part of a study to determine the relative intelligibility of products currently on the market. They were asked to listen carefully, and to identify each word that they heard by typing it on a computer-controlled keyboard when prompted on the computer screen. They were told that, if the speech did not sound like a real English word, or if it sounded like a word they did not know how to spell, they were to type in how they thought it would be spelled if it were a word in English. The instructions emphasized that subjects were to be as accurate as possible in making their identifications.

The experiment began with a short repetition of these instructions displayed on a computer screen. This was

followed by the presentation of the first trial in the first block. Immediately following the presentation of each word, subjects saw a prompt on their display screen to type the word they heard. At the beginning of each block, subjects heard a sample word presented in each of the four voices to familiarize them with the voices. This set of familiarization trials was not included in the set of test words in the trial, but it had as many syllables as the rest of the words in the trial.

Subjects' responses were analyzed by a trained phonetician to determine the number of words correctly identified. Only responses that were correctly spelled, were correct spellings of homophones of the intended stimulus word, or which could not be interpreted as being pronounced in any way other than the word intended by the talker were scored as correct. A percent-correct score for each condition was calculated. Because raw percentage scores do not fit the assumptions of standard analyses of variance (they are taken from a distribution bounded by 0 and 1), it was necessary to transform the scores such that the transformed scores more closely resembled samples from a normal distribution. For this study we used the standard arcsin transform (Kirk, 1995). Analyses were performed on both the raw and the transformed scores. ANOVA results (F values and estimates of probability) are reported using the more reliable transformed data, but graphs and difference scores were calculated using the untransformed percentages for ease of interpretation. For the purposes of analysis, the two male human talkers' scores were combined into a single human talker score.

Results

Figure 2 shows word recognition accuracy for the monomorphemic words for the three types of talker (Human, DEC and Votrax), expressed in terms of the percentage of all words in the block that were recognized correctly.

The percentage of spoken words that was identified correctly is plotted for monosyllabic words, bisyllabic words, and polysyllabic words. Overall, the results show a reliable difference in the intelligibility of the talkers ($F(2,24) = 121.209$, $p < .001$). The human talkers are 34 percentage points more intelligible than Votrax, ($F(1,24) = 606.661$, $p < .001$). In addition, the two human talkers are slightly, but reliably, more intelligible than DECTalk by eight percentage points ($F(1,24) = 41.250$, $p < .001$). Finally, DECTalk was

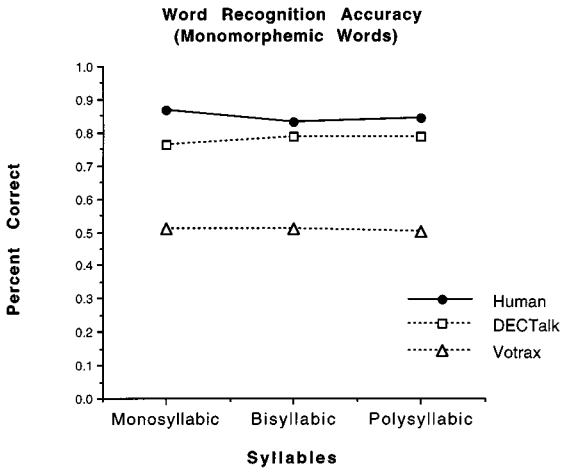

Figure 2. Word recognition accuracy by talker and number of syllables in word.

27 percentage points more intelligible than Votrax for all three word lengths for the monomorphemic words ($F(1,24) = 331.526$, $p < .001$). Thus, overall the two human talkers are more intelligible than DECTalk, and DECTalk is more intelligible than Votrax.

Although there was no effect of number of syllables on intelligibility for the monomorphemic words ($F(2,24) = .101$, n.s.), there was a significant interaction between number of syllables and the talker ($F(4,48) = 4.923$, $p < .01$). This suggests that the length of monomorphemic words does not affect their intelligibility overall, but the difference between listeners' ability to recognize longer words as compared to shorter words is in part dependent upon which voice produced the words. Bisyllabic monomorphemic words were three percentage points less intelligible than monosyllabic monomorphemic words when produced by human talkers, (from 87% correct for monosyllabic to 84% correct for bisyllabic, $F(1,48) = 7.142$, $p = .01$), which is consistent with the observation that more frequent or familiar words (the monosyllables) are more easily recognized than less frequent ones (see below). In contrast, bisyllabic monomorphemic words produced by DECTalk were four percentage points *more* recognizable than corresponding monosyllables (from 76% correct for monosyllabic to 80% correct for bisyllabic, $F(1,48) = 4.314$, $p < .05$). Finally, the 4-point difference between bisyllabic and monosyllabic monomorphemic words produced by Votrax was not significant (from 51% correct for monosyllabic to 55% correct for bisyllabic, $F(1,48) = 2.254$, n.s.). In contrast, polysyllabic words produced by all of the

voices were not significantly different in intelligibility from bisyllabic words. Recognition of human polysyllabic monomorphemic words was 87% (up three points from bisyllabic, $F(1,48) = 2.621$, n.s.), for DECTalk polysyllabic monomorphemic words it was 83% (up three points from bisyllabic, $F(1,48) = 1.077$, n.s.), and for Votrax polysyllabic monomorphemic words recognition was 51% (down four points from bisyllabic, n.s.).

These results suggest that word length can provide some small benefit when listening to synthetic speech. Although the bisyllabic monomorphemic words used in this study are significantly less recognizable than monosyllabic monomorphemic words when spoken by human talkers, they are significantly more recognizable when produced by DECTalk. This suggests that having more acoustic information about a talker's speech provides some assistance in interpreting that speech when it is impoverished (cf. Grosjean, 1985). The same pattern of results is observed for speech produced by Votrax, although the fact that this improvement is not significant suggests that the word-length advantage is small, and depends strongly on the acoustic-phonetic structure of the talker's speech being sufficiently human-like in the first place. The observation that polysyllabic words are not significantly more intelligible than bisyllabic words for any talker further suggests that any advantage conveyed by word length is not particularly strong.

In order to examine the effects of morphological complexity on intelligibility, we eliminated the monosyllabic items since, given the nature of the stimuli used, these words cannot vary on this dimension. Eliminating the monosyllabic words from analysis allowed us to compare the effects of morphological complexity and number of syllables on intelligibility for the four talkers using a standard repeated-measures ANOVA. As with the monomorphemic words, overall, the relative number of syllables in polymorphemic words did not reliably affect their intelligibility, ($F(1,12) = .009$, n.s.).

However, there were some apparent differences in the effect of morphological complexity across all talkers for the bisyllabic words and the polysyllabic words ($F(2,24) = 4.712$, $p = .05$). As a result, we will examine the degree to which morphological complexity affects the intelligibility of bisyllabic and polysyllabic words separately.

Figure 3 shows mean percent correct word identification for the bisyllabic words as a result of

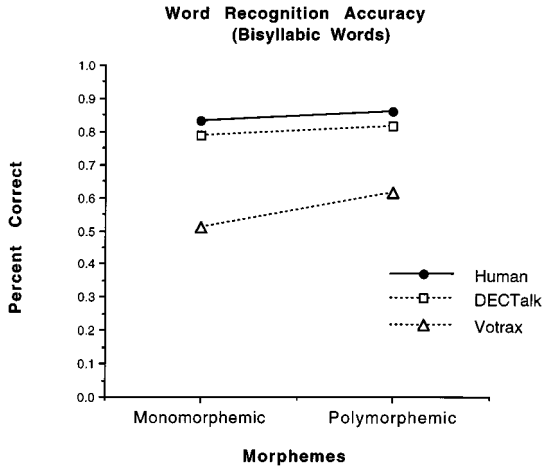

Figure 3. Word recognition accuracy for bisyllabic words, by talker and number of morphemes.

morphological complexity for each of the talkers. For the two-syllable words there was no reliable difference in recognition performance between the natural speech of the humans and the synthetic speech produced by DECTalk, ($F(1,24) = .267$, n.s.). However, the two human talkers were 28 percentage points more intelligible than Votrax ($F(1,24) = 97.587$, $p < .001$), and DECTalk was 24 percentage points more intelligible than Votrax ($F(1,24) = 67.403$, $p < .001$).

Furthermore, as can be seen in Figs. 3 and 4, across talkers there was an overall effect of morphological complexity ($F(1,12) = 4.712$, $p = .05$). Polymorphemic words were, on the whole, more easily

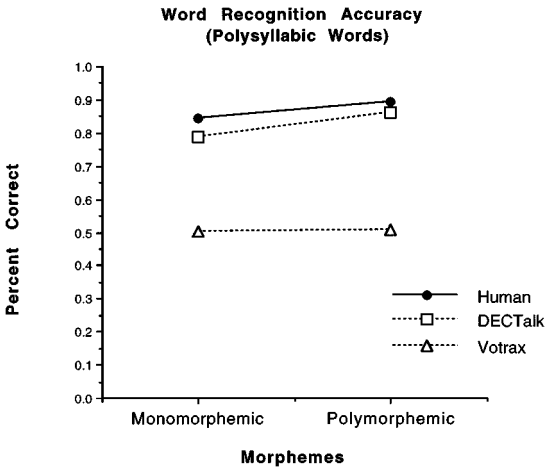

Figure 4. Word recognition accuracy for polysyllabic words, by talker and number of morphemes.

recognized than monomorphemic ones. Looking at individual talkers, for bisyllabic words there was an increase of three percentage points for words produced by human talkers (from 83% to 86%), and a 3-point increase for words produced by DECTalk (from 79% to 82%), though these differences were not significant ($F(1,24) = .698$, n.s., ($F(1,24) = .011$, n.s., respectively). By comparison, for Votrax-produced speech, two-morpheme words were reliably more intelligible than one-morpheme words, ($F(1,24) = 4.407$, $p < .05$), showing an increase of 11 percentage points, from 51% to 62%. Thus, for the two-syllable words, morphological complexity aided word recognition, at least in the case of the least intelligible synthetic speech. This suggests that listeners are able to use their knowledge of morphological complexity to aid in recognition, but this knowledge is most useful when the speech is otherwise difficult to recognize.

A somewhat different pattern of results is seen in the effects of morphological complexity across the different talkers when we examine recognition of three- and four-syllable words, as shown in Fig. 4. For natural speech and for synthetic speech produced by DECTalk, increasing morphological complexity resulted in a slight increase in the intelligibility of the speech by five percentage points for natural speech and seven percentage points for DECTalk speech, but these differences are not significant ($F(1,24) = 2.907$, $p = .1$; $F(1,24) = 3.921$, $p = .06$, respectively). Increasing the morphological complexity of three and four-syllable words increases intelligibility by only one percentage point (from 50% to 51% correct) for the Votrax-produced synthetic speech, and this difference is similarly not significant ($F(1,24) = .002$, n.s.).

Overall, a greater degree of morphological complexity significantly aided recognition. For the longest words this difference was not significant for any individual voice. However, when listening to shorter words, increasing morphological complexity provided a benefit to recognition most clearly for words produced by the least intelligible speech synthesizer (Votrax). This suggests that though listeners may be able to use their knowledge of words' morphological structures to facilitate recognition, this ability is most useful when acoustic information about the segmental characteristics of the word is impoverished. However, as acoustically difficult words get longer, this advantage is lost. One reason for this may be that, impressionistically, it is very easy to "lose" the speech-like quality of the Votrax speech—the longer a given passage is, the more likely it is that the Votrax speech momentarily stops sounding like speech at all. As words get longer (from two syllables to three or four syllables), the benefit provided by complex morphology disappears as words are more likely to become unrecognizable as speech.

In studying intelligibility it is important to control for factors such as the familiarity or frequency of the words to be identified, as these parameters have been shown to be highly significant in word recognition (Kucera and Francis, 1967). The words used in this study were not matched a priori for frequency of occurrence or for familiarity since our word-frequency lists did not include enough words that otherwise fit the morphological, syllabic, and overall phonological requirements of the present study (cf. Cutler, 1981). However, we did perform an analysis of those words in our list for which we have frequency (Kucera and Francis, 1967) and familiarity (Nusbaum et al., 1984) data (40% of our words were found in Kucera and Francis, while 68% were found in Nusbaum et al.). The results of this analysis show that the intelligibility results discussed above for synthetic speech are not a result of the effects of either frequency or familiarity (although, as mentioned previously, the results for monomorphemic words produced by the human talkers can be explained in terms of word frequency and familiarity).

Figure 5 shows that, among the multisyllabic words, polymorphemic words are considerably less frequent than monomorphemic words. Though this difference is not significant ($F(1,22) = .958$, n.s.), the trend is clear. If the observed differences in the intelligibility

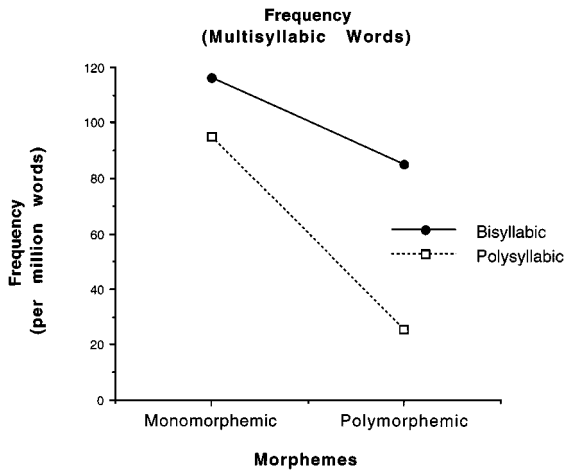

Figure 5. Frequency of multisyllabic (bisyllabic and polysyllabic words) by number of morphemes.

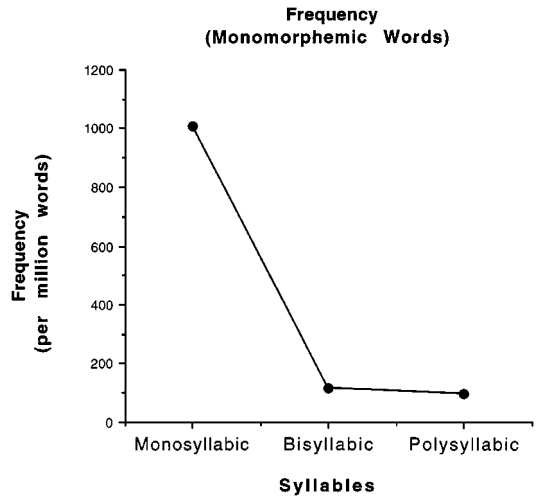

Figure 6. Frequency of monomorphemic words.

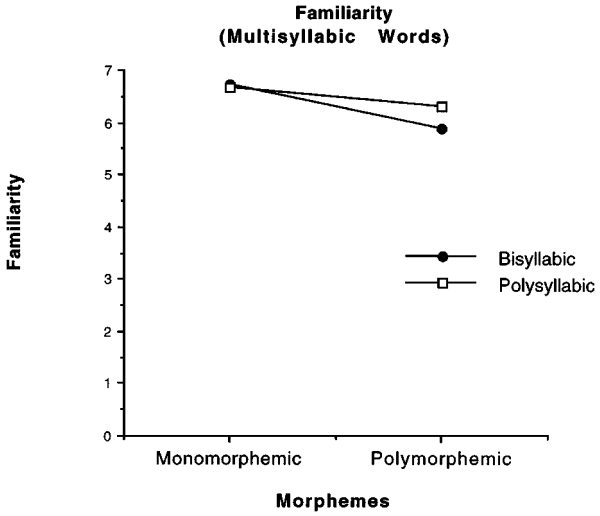

Figure 7. Familiarity of multisyllabic (bisyllabic and polysyllabic words) by number of morphemes.

of words produced by DECTalk and Votrax were due entirely to word frequency effects, these differences would predict that polymorphemic words should be *less* intelligible than monomorphemic words. However, our results show that the polymorphemic words were in general more intelligible than monomorphemic words. This means that any improved intelligibility of polymorphemic words cannot be explained by their higher frequency of occurrence in English.

This pattern of more frequent words having a measurably lower intelligibility rating holds even for the monomorphemic words, as shown in Fig. 6. In this case, monosyllabic words are clearly more frequent than either bisyllabic or polysyllabic words, though because of the huge degree of variability the difference is not significant ($F(1,24) = 3.105$, n.s.). However, despite the large difference in frequency, monosyllabic, monomorphemic words are still not as intelligible as bi- or polysyllabic monomorphemic words when produced by speech synthesizers.

Similar results for the effect of familiarity may be seen in Fig. 7. The familiarity scores were collected by having undergraduate students at Indiana University rate how familiar they are with each word in a large set (Nusbaum et al., 1984). As with frequency, although the difference in familiarity between the monomorphemic and polymorphemic words was not significant ($F(1,38) = 3.815$, n.s.) the trend was toward polymorphemic words being less familiar than monomorphemic words. This would predict that polymorphemic words should be less intelligible than monomorphemic words, which was not the case.

The identification performance for the monomorphemic words showed the same lack of familiarity effects on intelligibility. Figure 8 shows that there is a slight trend toward lower familiarity scores for words with larger numbers of syllables (monosyllabic ≫

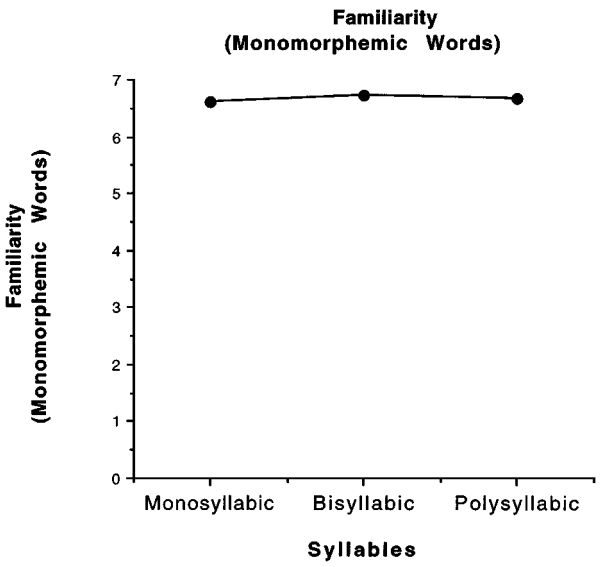

Figure 8. Familiarity of monomorphemic words.

bisyllabic $\gg$ polysyllabic). As before, this difference was not significant ($F(1,36) = .076$, n.s.), but the trend of familiarity was again opposite what one would normally expect from the results of the recognition of synthetic speech. In this case, more familiar words were generally less intelligible.

Thus, though the words used in this study did differ in terms of their frequency and familiarity, the results observed for synthetic speech cannot be explained in terms of the standard finding that more frequent or more familiar words are more intelligible than less frequent or less familiar ones (e.g., Marlsen-Wilson, 1987).

In summary, there were three general findings in this experiment. Word length alone did not significantly aid word recognition, although for monomorphemic words produced with higher quality synthetic speech there was some benefit to increasing word length from monosyllables to bisyllables. Furthermore, morphological complexity did affect word recognition overall, although the strength of the effect depended on word length and the quality of the speech. For two-syllable words with high-quality speech, morphological complexity did not significantly affect word recognition, possibly because there was sufficiently high quality phonetic information to permit identification on the basis of acoustic patterns alone. By contrast, for two-syllable words produced by a less intelligible synthesizer from which phonetic information alone is likely too poor to allow recognition, listeners were able to use their lexical knowledge about possible morphological combinations to improve their recognition of two-morpheme bisyllabic words above their recognition of monomorphemic bisyllabic words. The converse was observed for three- and four-syllable words. At this word length, morphological complexity did not aid in recognizing words produced using the least intelligible speech synthesizer (perhaps because the poverty of phonetic information was greater due to the greater word length), but it did somewhat improve word recognition for speech produced by humans and a more intelligible synthesizer.

Discussion

In understanding speech, listeners use information from many domains. Bottom-up information about the segmental and prosodic characteristics of the speech signal is supplemented with top-down information about morphological structures (among other things) (Marlsen-Wilson and Welsh, 1978; Grosjean and Gee, 1987). However, in order to make most efficient use of higher level knowledge, listeners must be able to interpret a sufficient amount of the acoustic information that comprises the word (cf. McClelland and Elman's 1986 TRACE model). If the acoustic cues of a word are not intelligible or are misleading in terms of the segmental or prosodic patterns they indicate, then listeners may not be able to make use of their knowledge of morphological structure to improve recognition. Thus, speech synthesizers that are accurate at reproducing the acoustic cues of natural speech further facilitate recognition by allowing listeners to use their full range of knowledge of the patterns of spoken language.

Generally speaking, listeners recognized polymorphemic words more accurately than monomorphemic words. Even though polymorphemic words tend to be slightly less familiar and less common, listeners can clearly use the structural constraints provided by morphological structure to aid in word recognition. However, when the segmental structure of speech is difficult to recognize, as for speech produced by the Votrax text-to-speech system, morphological complexity only aids recognition for bisyllabic words. Apparently, the three- and four-syllable words produced by Votrax are sufficiently difficult to recognize that listeners still cannot make effective use of the morphological constraints that reduce the number of possible interpretations of longer words. For higher quality speech, recognition performance is relatively good for the two-syllable words so morphological complexity does not provide much assistance, and little benefit is observed. But with longer words, listeners can more effectively use morphological structure to aid in recognition. One possible explanation for this is that the DECTalk synthesizer more accurately reproduces those natural speech phenomena that provide cues to the more complex structures of longer words.

These results suggest that intelligibility tests need to measure segmental intelligibility in a wider range of linguistic contexts than the monosyllabic words currently used. Such tests will be useful for developers of systems incorporating synthetic speech as well as the developers of speech synthesis systems themselves. Clearly, the availability of the results of more detailed diagnostic intelligibility tests will allow users

of synthetic speech to make more informed choices about which systems to use for particular applications. In particular, knowing the limitations of a synthesis system can aid in designing applications that make use of the strengths of the synthesizer while minimizing the effects of its weakness. From the perspective of the development of better speech synthesis systems, intelligibility tests which incorporate multisyllabic and multimorphemic words will provide diagnostic information for improving the acoustic-phonetic rules in contexts that are not present in monosyllabic words (cf. Spiegel et al., 1990). Finally, such tests are clearly vital for developing better rules which replicate the effects of morphological complexity in natural speech. The information provided by the current standard tests of monosyllabic, monomorphemic words cannot diagnose the way text-to-speech systems encode morphological complexity into speech. Because, as we have demonstrated here, this complexity does affect intelligibility, it is important to design new intelligibility tests that permit assessments that are more diagnostic of performance with the kinds of messages and linguistic materials that may be used in real-world applications.

Appendix A

	Monosyllabic	Bisyllabic	Polysyllabic
Monomorphemic	case	able	anchovy
	change	donut	dolomite
	claim	employ	enamel
	do	enter	examine
	dole	exit	guillotine
	gym	gymnast	gymnasium
	imp	gypsum	imperial
	let	imply	personal
	place	migrate	property
	stress	patient	proportion
	vest	person	strategy
		portion	sycamore
		proclaim	vesicle
		promise	
		streusel	
		subtle	
		versus	
Polymorphemic		abler	disabler
		disable	dolefully
		doable	emplacement
		doleful	encasement
		emplace	exchangeable
		encase	immigration
		exchange	impatient
		impish	impatiently
		placement	impishly
		stressful	migration
		sublet	patiently
		undo	personality
		vestment	proclamation
			proportionate
			stressfully
			subletting
			undoable
			vestmented

Acknowledgments

This is the text of a paper that was presented at the 131st meeting of the Acoustical Society of America, May, 1996, Indianapolis. This research was supported in part by a grant from the Digital Equipment Corporation and in part by a grant from the Social Sciences Division of the University of Chicago. We thank Tony Vitale for numerous discussions on the issues raised in this paper, and Dr. Daryle Gardner-Bonneau and two anonymous reviewers for their comments and suggestions.

References

Cole, R.A. and Rudnicky, A.I. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874–1946. *Psychological Review*, *90*:94–101.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*:65–70.

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception & Psychophysics*, *28*:267–283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics*, *38*:299–310.

Grosjean, F. and Gee, J.P. (1987). Prosodic structure and spoken word recognition. *Cognition*, *25*:135–156.

Kirk, R.E. (1995). *Experimental Design*. Pacific Grove: Brooks/Cole Publishing Co.

Kucera, H. and Francis, W.N. (1967). *Computational Analysis of Present-day American English*. Providence: Brown University Press.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*:2018–2024.

Logan, J.S., Greene, B.G., and 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., Feustel, T.C., and Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors*, *83*:17–32.

Marlsen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*:135–156.

Marlsen-Wilson, W.D. and Welsh, A. (1978). Processing interactions during word-recognition in continuous speech. *Cognitive Psychology*, *10*:29–63.

McClelland, J.L. and Elman, J.L. (1986). The TRACE model of speech perception. In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: The MIT Press, pp. 58–121.

Nusbaum, H.C. and Pisoni, D.B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, *17*:235–242.

Nusbaum, H.C., Pisoni, D.B., and Davis, C. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. Research on Speech Perception Progress Report No. 10, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, IN.

Ralston, J.V., Pisoni, D.B., and Mullennix, J.W. (1995). Perception and comprehension of speech. In A.K. Syrdal, R.W. Bennett, and S.L. Greenspan (Eds.), *Applied Speech Technology*. Boca Raton, FL: CRC Press, pp. 233–288.

Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*:474–494.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A.K. Syrdal, R.W. Bennett, and S.L. Greenspan (Eds.), *Applied Speech Technology*. Boca Raton, FL: CRC Press, pp. 195–232.

Slowiaczek, L.M. and Nusbaum, H.C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, *27*:701–712.

Spiegel, M.F., Altom, M.J., Macchi, M., and Wallace, K.L. (1990). Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication*, *9*:279–291.