cognitive load, see Table VI. There are only a few studies that have attempted quantitative evaluations of text-to-speech systems to date; much of the data on the capabilities and limitations of the current technology comes from work performed at Indiana University by David Pisoni and his colleagues (Pisoni *et al.*, 1985).

A. Intelligibility of isolated words

The measurement of intelligibility can be performed in many different ways. Since consonants have been more difficult to synthesize than vowels, the modified rhyme test (House *et al.*, 1965) is often used, in which the listener selects among six familiar words that differ only by an initial consonant or a final consonant. This is not a very severe test of system performance since the response alternatives may exclude a confusion that would be made if a blank answer sheet were used, but the test does facilitate rapid presentation to naive subjects and automatic scoring of answer sheets.[15] If possible, an open response, including perhaps a rating of goodness of each item, should be used with such a test in order to better determine systematic error patterns and deficiencies, especially if there are relatively few errors.

Logan *et al.* (1986) evaluated the intelligibility of eight text-to-speech systems by presenting listeners with a recording of the modified rhyme test words. The results are summarized in Table VII. Also included are comparable data obtained earlier with the Haskins text-to-speech system (Cooper *et al.*, 1984). Systems are rank ordered according to performance. When percent correct is fairly high, a good way to compare systems is to use percent error (simply 100 minus percent correct) because relative changes in percent error better reflect the difficulty of comprehension and the difficulty of making improvements. The frequency of occurrence of perceptual errors in running text is approximated by the reciprocal of the percent error values given in the table.

TABLE VI. Techniques for evaluating text-to-speech systems.

INTELLIGIBILITY:	Diagnostic rhyme test (Fairbanks, 1958; Voiers, 1983)
	Modified rhyme test (House *et al.*, 1965)
	Open response rhyme test (Pisoni *et al.*, 1985)
	MRT in noise (Nixon *et al.*, 1985)
	CNC word list (Lehiste and Peterson, 1959)
	CVC nonsense words (Dubno and Levitt, 1981; Pols and Olive, 1983)
	CID W-22 word list (Hirsh *et al.*, 1952)
	Goodness ratings for words (Wright, Altom, and Olive, 1986)
	CID sentences (Erber, 1979)
	Harvard sentences (Egan, 1948)
	SPIN test (Nakatani and Dukes, 1973; Kalikow *et al.*, 1977)
	Haskins anomalous sentences (Nye and Gaitenby, 1974)
COMPREHENSION:	Reading/listening comprehension (Pisoni and Hunnicutt, 1980)
	Sentence verification (Manous *et al.*, 1985)
NATURALNESS:	Paired comparisons (IEEE, 1969; Logan and Pisoni, 1986)
	Subjective ratings (Nusbaum *et al.*, 1984)

TABLE VII. Performance of selected text-to-speech systems with respect to CVC intelligibility using the modified rhyme test, closed response, after Logan *et al.* (1986) and Cooper *et al.* (1984).

Device	% correct	% error
Type-n-Talk	73	27
Infovox	88	12
MITalk-79	93	7
Prose-2000 3.0	94	6
DECtalk 1.8	97	3
Natural speech	99	1
Haskins system	93	7
Natural speech	98	2

Looked at in this way, the expected rate of perceptual errors for DECtalk is about (100%/3%), or one segmental misperception about every 33 syllables of text. The error rate for the Prose-2000 is about twice that of DECtalk, while it appears that Type-n-Talk is seriously flawed (see also Cochran, 1986).

When Logan *et al.* (1986) ran the same vocabulary used in the modified rhyme test, but with open response, the error rate went up quite a bit—typically 3 to 4 times the closed-response error rate—but the relative rankings of systems did not change. Open response, however, had the advantage that systematic error tendencies could be detected and (hopefully) corrected. For example, DECtalk 1.8 had a problem with nasals adjacent to high front vowels—a problem that was then corrected in DECtalk 3.0. The test used is perhaps not ideal for detection of all likely consonantal confusions because the words are not particularly well balanced phonetically, and there are no consonant clusters or unstressed syllables. Other word lists address some of these deficiencies (Lehiste and Peterson, 1959; Nusbaum *et al.*, 1984), but there is a clear need for better diagnostic instruments in the evaluation of text-to-speech systems.

The intelligibility of several linear prediction based systems has been studied by Pols and Olive (1983). They presented consonant–vowel–consonant (CVC) nonsense syllables to high school students after a brief introduction to phonemic representations. The syllables were either (1) natural speech digitized at 10 000 12-bit samples/s, (2) 10-pole linear-prediction coded versions of these syllables, or (3) syllables synthesized using the Olive (1977) LP diphone concatenation scheme. The results are shown in Table VIII. This is a very difficult task for naive unpracticed subjects, as indicated by the relatively low 93% phoneme recognition performance for natural speech.[16] Two points of interest are that (1) linear prediction coded speech can suffer a serious reduction in intelligibility, even when there is no effort to

TABLE VIII. Consonant intelligibility in nonsense syllables encoded in various ways (Pols and Olive, 1979).

Condition	% correct	Typical errors
OLIVE (1977) DIPHONE SYNTHESIS	66	voicing, nasality
LPC-10, no quantization	86	b-v-ð,m-n-ŋ
DIGITIZED NATURAL, 5 kHz, 12 bit	93	f-θ,v-ð

save bits by quantizing the representation, and thus (2) linear-prediction coded speech often gives listeners more favorable impressions of intelligibility and naturalness than are warranted by objective measures.

Based on this critical evaluation, Olive went on to select new versions of his diphone inventory, also hand-correcting pitch errors, and retested diphone intelligibility iteratively until the most recent system exceeds the intelligibility of LP-10. Part of the intelligibility increase may be attributed to the use of multipulse linear prediction (Atal and Remde, 1982; Olive and Liberman, 1985), which makes possible the detailed modeling of bursts of noise and other syllable-onset events.[17]

Wright *et al.* (1986) discovered that it is possible to detect deficiencies in segmental synthesis even when intelligibility is relatively high, simply by asking subjects to rate the subjective goodness of words. Naive listeners hear a word and then see a visual presentation of the word, at which point they are asked to rate goodness. If the goodness rating is low, the computer asks additional questions about the location and type of specific defects.

In an effort to find a maximally sensitive test for comparing phoneme intelligibility of various systems, Nixon *et al.* (1985) added controlled amounts of background noise to synthesized or vocoded MRT word lists, and measured intelligibility as a function of signal-to-noise level. They found that an unidentified "high-performance" text-to-speech system was about six percentage points worse than natural speech over a wide range of S/N ratios. Stated in another way, under adverse S/N conditions, the synthetic speech had to have a 5-dB boost in S/N ratio to be as intelligible as natural speech.[18] Of perhaps greater interest are comparative figures from the Nixon study for 2.4-kbit government-standard LPC-10, and 9600-bit CVSD, both of which performed much worse than the synthetic speech produced by this text-to-speech system—both being about 40% less intelligible than natural speech at high and low S/N ratios. These rather surprising results suggest limits to the utility of low-bit-rate encoded speech, and suggest that, at least for some applications, text-to-speech systems already offer superior communicative performance.

B. Intelligibility of words in sentences

In comparison with words spoken in isolation, words in sentences undergo significant coarticulation across word boundaries, phonetic simplifications, reduction of unstressed syllables, and prosodic modifications that, among other things, shorten nonfinal syllables and modify the fundamental frequency contour. In order to evaluate the ability of text-to-speech systems to realize these transformations, tests of word intelligibility in sentence frames have been devised. The easiest materials, consisting of simple short predictable sentences known as the CID sentences (Erber, 1979), have been used primarily to evaluate abilities of the hearing impaired. Another sentence list was devised to measure speech intelligibility in noise (Egan, 1948). This list, known as the Harvard sentences, is often employed today, in spite of its meager syntactic variation and minimal use of words with more than two syllables, simply because no better lists have been proposed and calibrated. Pisoni *et al.* (1985) employed a subset of the Harvard Sentences and measured the intelligibility of each content word. The results are presented in Table IX. The same rank order of systems holds as was obtained for isolated words. Also shown in the table are data from a Haskins anomalous sentence test (Nye and Gaitenby, 1974), consisting of nonsensical word strings that were syntactically acceptable—of the form "The (adjective) (noun) (verb) the (noun)," e.g., "The old farm cost the blood." Again, system rank ordering is the same, but differences between systems are somewhat greater, suggesting that this is a more sensitive test.

The performance of the Haskins system, as evaluated by Ingemann (1978) and reported by Cooper *et al.* (1984) is also shown in the table. The poorer general performance of subjects in the Ingemann study on the natural speech control may imply that the scores should be boosted slightly before comparison with the systems listed above in the table.

Chial (1985) used the SPIN (speech in noise) test developed by Kalikow *et al.* (1977) and calibrated by Bilger *et al.* (1984) to evaluate the relative performance of several of the less expensive text-to-speech systems. Subjects had to identify the last word in sentences presented in a background babble of several competing voices. Included were the Echo II, the Votrax Type-n-Talk that incorporates the SC-01 synthesis-by-rule chip, and the Votrax Personal Speech System, which uses a new version of the chip, the SC-01A. Results, shown in Table X, indicate that the new chip has improved intelligibility over the SC-01, up from 40% to 65% words correct as measured at 0-dB signal-to-babble level. However, performance with natural speech at this signal-to-babble level is typically about 91% correct (Chial, 1985; Bilger *et al.*, 1984), so one must conclude that these inexpensive devices are still very limited in intelligibility.

C. Reading comprehension

Since synthetic speech is less intelligible than natural speech, what happens when one tries to understand long paragraphs? Do listeners miss important information? Is a listener so preoccupied with decoding individual words that the message is quickly forgotten? In an attempt to answer these questions, Pisoni and Hunnicutt (1980) included a standard reading comprehension task in their evaluations. Half the subjects read the paragraphs by eye, while the other half listened to a text-to-speech system. In a later experiment, comparison was made with a human voice reading the

TABLE IX. Performance of selected text-to-speech systems with respect to word intelligibility in Harvard test sentences and Haskins anomalous sentences, after Pisoni *et al.* (1985) and Cooper *et al.* (1984).

Device	Meaningful % correct	Anomalous % correct
Prose-2000	84	65
MITalk-79	93	79
DECtalk	95	87
Natural speech	99	98
Haskins system		78
Natural speech		95

TABLE X. Performance of inexpensive text-to-speech systems with respect to word intelligibility in the SPIN (speech in noise) test sentences, after Chial, 1985.

Device	% correct
ECHO-II	18
VOTRAX TYPE-N-TALK	40
VOTRAX PERSONAL SPEECH SYSTEM	65
Natural Speech	91

paragraphs. Results of answering multiple-choice questions about the content of the paragraphs are shown in Table XI. The text-to-speech systems performed about equally well, suggesting that the test is not sensitive enough to compare systems, and that the limit on performance is the memory capacity of these college students rather than the difficulty of comprehending synthetic speech. Pisoni also observed that subjects typically got better on the second half of the test when listening to synthetic speech, even though there was no feedback of correct answers.[19] On the second half, listening subjects performed about as well as the readers.

One might conclude that current text-to-speech systems produce quite satisfactory speech since there is no measurable decrement in listening comprehension after a familiarization period. Thus synthetic speech should be a viable method of presenting information over an auditory channel in most applications. Such conclusions are perhaps premature because (1) similar experiments have not been performed over the telephone, or with less-educated subjects, and (2) multiple-choice tests and recall measures may not be sensitive enough to reveal differences in perceptual processing between natural and synthetic speech. Pisoni (1982) used a reaction-time experiment to show that listeners do indeed devote somewhat more time to speech perception when exposed to synthetic speech as compared with natural speech, and Manous *et al.* (1985) measured a decrement in accuracy and speed of response for text-to-speech systems versus natural speech using a more sensitive comprehension test in which listeners had to immediately respond "true" or "false" to each sentence they heard. The capacity of short-term memory for earlier items in a list can also be reduced when listening to synthetic speech (Luce *et al.*, 1983).

In summary, studies have shown that there is a wide range of performance between text-to-speech systems in terms of segmental intelligibility. Measured in terms of error rate, a system with a 3% error rate is twice as good as one with a 6% error rate, at least in terms of the average time interval between misperceptions in running text. Language is sufficiently redundant that these differences in segmental intelligibility often appear to be slight, but this is not the case when listening to unfamiliar names or difficult material. Furthermore, errors are usually the result of deviations of synthesizer parameters from values seen in natural speech. To the extent that error rate reflects a tendency for misspecification of parameters in general, it is also an indicator of how *unnatural* the speech is likely to sound.

TABLE XI. Performance of several text-to-speech systems with respect to listening comprehension (percent of questions about paragraph contents that were answered correctly), compared with visual presentation, after Pisoni and Hunnicutt, 1980).

Device	% correct	
Natural speech	68	
MITalk-79	70	(75% on second half of test)
Prose-2000	65	
Visual presentation	77	

D. Naturalness

Naturalness is a multi-dimensional subjective attribute that is not easy to quantify. Any of a large number of possible deficiencies can cause synthetic speech to sound unnatural to varying degrees. Fortunately, systems can be *compared* for relative subjective naturalness with a high degree of inter-subject and test–retest agreement (IEEE, 1969; Munson and Karlin, 1962). A standard procedure is to play pairs of test sentences synthesized by each system to be compared, and obtain judgments of preference (Logan and Pisoni, 1986). As long as the sentences being compared are the same, and the sentences are played without a long wait in between, valid data can be obtained. It is more difficult to compare systems that have been heard on different days or with different synthetic materials since extraneous factors can add an unpredictable amount of "noise" into listener preference judgment data (Nusbaum *et al.*, 1984).

Naturalness should not be confused with intelligibility. Some of the low bit rate linear-prediction systems sound like slightly distorted recordings of natural speech (which is what they are), and so are judged fairly natural, but they test out to have rather poor intelligibility scores (Nixon *et al.*, 1985). On the other hand, intelligibility and naturalness ratings of text-to-speech systems appear to be fairly highly correlated.

E. Suitability for a particular application

Text-to-speech devices are being introduced in a wide range of applications. A sampling of commercial uses appears in Table XII. Noncommercial applications are described in Sec. V. These devices are not good enough to fully replace a human, but they are likely to be well received by the general public if they are part of an application that offers a new service, or provides direct access to information stored on a computer, or permits easier or cheaper access to a present service because more telephone lines can be handled at a given cost. Both intelligibility and naturalness are considered important factors to the success of any application, but it is interesting to note that one large commercial concern is planning an application that will use DECtalk set up to speak in a monotone, purposely trying to indicate to the customer that he/she is talking to a smart computer rather than to a poor imitation of a human. What is important at this early stage in the exposure of the public to synthetic speech is to avoid applications that might lead to user frustration and generate negative attitudes toward all devices that "*talk like a computer.*" For example, intelligibility over the telephone

TABLE XII. Selected commercial applications for text-to-speech.

TEXT-TO-SPEECH BUSINESS APPLICATIONS
• Telephone information: e.g., 800 numbers for stock quotations, weather, ski conditions, sports scores, museum exhibits/schedules, talking Yellow Pages, ... (information that is changed frequently, and is available in computerized text form)
• Remote (on the road) access to computer mail
• Catalog ordering by phone, banking by phone (requires keypad or speech recognition for input)
• Data-base inquiry, especially for unsophisticated users: e.g., sales reps can determine status of purchase orders
• Generation of cassette recorded instructions for assembly plants, backplane wiring, telephone circuits, etc. (Flanagan *et al.*, 1972)
• Telephone access to computerized repair "experts" on, e.g., computers, telephone circuits.
• Coordination of large numbers of people on the road through a central computer information bank
• Warning and alarm systems concerning malfunctioning equipment
• Talking terminals and training devices (speech is often better than reading)
• Proofreading (catches kinds of typing errors that are often hard to detect visually)

of current text-to-speech systems may not be adequate for applications that involve many unfamiliar names (the telephone itself is somewhat marginal for this purpose, synthesis implies a further intelligibility reduction, and relatively poorer performance of these systems in converting names to phonemes adds additional potential confusion).

A current limitation for systems using the telephone is that the computer cannot listen as well as it can talk. Speech recognition technology is lagging behind synthesis capabilities. Presently, any user responses must be entered by telephone key pad commands, and not all telephones are push-button phones. Speaker-independent connected digit recognition systems are now being demonstrated in the laboratory with better than 98% string recognition accuracy (Bush and Kopek, 1986); perhaps this technology will become commercially available soon.

Eventually, text-to-speech systems will compete with products now used to produce canned messages from waveform-coding chips. Currently, such waveform encoding systems are thought to produce far more natural speech (due to natural timing, intonation, and voice quality obtained from a human utterance), even though measured intelligibility is significantly lower than for the better text-to-speech systems (Nixon *et al.*, 1985). If text-to-speech systems can be increased in naturalness even slightly, the advantages in terms of ease of message assembly can easily outweigh the cost advantage accruing to the waveform coder for many applications.

V. SPECIAL APPLICATIONS

Text-to-speech systems are beginning to be applied in many ways, including aids for the handicapped, medical aids, and teaching aids. This brief section is included in the hope of stimulating additional humanitarian applications for this new technology. It is an unfortunate fact of life in the United States that funds for transfer of this technology to the handicapped are scarce, and funds to actually purchase devices for individuals are virtually nonexistent. We depend on the success of text-to-speech systems in the commercial marketplace to lead to less costly portable low-power devices that, through the good will of industry, may be made available at special discounts for the handicapped.[20]

A. Talking aids for the vocally handicapped

The first kind of aid to be considered is a talking aid for the vocally handicapped. There are over 1.5 million nonspeaking persons in the USA, excluding the deaf, according to a survey made by the American Speech and Hearing Association (ASHA, 1981). Any person in this group who can point at some kind of a communication board or use a typewriter keyboard is a potential user of a communication aid that involves conversion of text to speech. A continually updated listing of communication aids for the nonvocal is maintained by the Trace Center of the University of Wisconsin (Vanderheiden, 1978, 1985); see also the quarterly publication *Communication Outlook* (Portnoy, 1979–present) and Bernstein (in press).

A potential advantage of DECtalk in this application is the possibility of fitting the voice characteristics to the user, particularly the advantage of giving women a femalelike voice and children a childlike voice. Prior to the availability of DECtalk, a 16-year old girl in Arizona who was injured in an automobile accident refused to use a talking aid because it made her sound masculine. On the other hand, some young cerebral palsy children seem to enjoy having a robotlike monotone voice speak for them when among their peers in a classroom setting.

Warrick *et al.* (1977) identify a number of capabilities that would facilitate wider use of talking aids: (1) natural distinguishable voices for each child in a classroom, (2) ability to express emphasis and attitude, (3) lighter weight and more portable configurations, (4) predictive type-ahead or other methods for speeding text specification. As pointed out by Bernstein (1986), natural voices are distinguished from one another by many types of cues that not only signal gender, but also approximate size, age, and regional accent. Current synthesis algorithms modify vocal tract size and laryngeal waveform to distinguish among a small set of speakers, but do not include capabilities to modify dialect, timing, intonation, allophonic selection, or phonetic realization. Users of talking aids can be frustrated by an inability to convey emotions such as urgency or friendliness by voice. Everything comes out in a sort of semantically neutral way, although some systems provide an ability to emphasize selected words.

The vocally handicapped present a wide range of motor difficulties that requires ingenious solutions to permit text creation. One method for speeding up text input is to use a predictive input system that always displays the most frequent English word for any typed word fragment, and the user can hit a special key to accept the prediction (Hunnicutt, 1985). Another alternative, similar in some ways to shorthand, is the Bliss symbol system (Carlson *et al.*, 1982b;

Hunnicutt, 1984). Each symbol stands for a common word. The authors found that Bliss symbols seem to be a good way to get nonvocal children started on language production, but a switch to normal orthography seems desirable later.

Vocal communication difficulties are fairly common when many prelingual deaf try to communicate with a hearing individual who is unfamiliar with the speech of the deaf. Bernstein *et al.* (1984) have designed a telephone communication system to overcome this problem. A deaf person can speak utilizing a Speech Plus text-to-speech board, and can "listen" by viewing the output of a large-vocabulary isolated-word speech recognition device. Preliminary data suggest that the performance of the recognition system in current use is marginal, but still good enough to be useful; future improvements could make such systems more attractive. An important issue is how to format the recognition alternatives, e.g., phonemes or a word hypothesis lattice (Huggins *et al.*, 1986).

B. Training aids

In one sense, a talking aid is by default a language training aid because it promotes practice and elicits direct feedback. This is one reason why it would be advantageous to get more talking aids to nonvocal children as early as possible in their school career. Experience suggests that this kind of device will also promote correct spelling and syntax (Carlson *et al.*, 1980). The inherent attraction of computer devices may mean that the approach could also be used with normal children for initial reading instruction.

A novel and quite successful application of text-to-speech is in the area of training dyslexic children to read. Dyslexia is a self-perpetuating difficulty because it is embarrassing to be helped by a teacher or friend, and it is nearly impossible to practice reading without help. Now several research groups have devised computer systems that permit unsupervised reading practice (Atkinson, 1972; Olson *et al.*, 1985). For example, the system being developed at the University of Colorado–Boulder (Olson *et al.*, 1985), uses a computer display screen, a mouse pointer, and a DECtalk text-to-speech system to read unfamiliar words or sound them out syllable by syllable.

Training aids need not be restricted to handicapped individuals (Sherwood, 1981). It is well known that speech has measurable advantages over reading and writing in many cognitive situations (Ochsman and Chapanis, 1974). For example, Suppes (1979, 1981) devised a computerized course for teaching algebra, and showed that providing some of the interactions via spoken responses resulted in better learning performance than visual presentation of all computer responses. Nakatani *et al.* (1986) provide spoken tutoring in the use of a text editor, noting that otherwise the student must constantly switch attention between the behavior of the editor and any tutorial information provided at the bottom of the screen. Tutorials of this sort can be made to have quite natural intonation and phrasing by proper annotation of the tutorial text (Hirshberg and Pierrehumbert, 1986).

C. Reading aids for the blind

Another application area is the development of reading aids for the blind. According to a survey by the National Center for Health Statistics (NCHS, 1977), approximately 1.4 million Americans are so severely visually impaired that they are unable to read ordinary news print, even with glasses. Machines that can scan printed material and produce speech would be of great help to this community. Ultimately, the goal is personal reading machines, although the current cost of the best performing machine, Kurzweil's, at a price of over $30 000, is far from this objective.

With the advent of computer typesetting, and the large text data bases available to computer users, it may not be necessary for a blind person to obtain a high-cost text reader. In some cases, connecting a text-to-speech system to a personal computer that is interfaced to a large network may serve many information gathering needs. One interesting pilot project in Sweden uses an FM overnight broadcast to load the day's newspaper into a blind person's personal computer, and has indexing programs to permit scanning of topics, using the Infovox SA-101 text-to-speech system (Carlson *et al.*, 1976; Carlson *et al.*, 1981).

Other efforts in this area have been concerned with timely production of talking books for the blind. Currently, most cassette recordings of books for the blind are produced by Recording for the Blind in Princeton, New Jersey. They use volunteer readers, and find that it takes up to 6 months after an order is placed to produce an audio copy of a typical textbook—primarily because of the problems inherent in coordinating volunteer efforts when so many hours of speech are to be produced. There are two possible ways to speed up the delivery of textbook orders. A text-to-speech system can work day and night to produce audio cassette tapes, or the text could be placed on a disk for a personal computer—in which case a blind person having a personal text-to-speech system could listen to the book and potentially be able to skim and scan over the book much more efficiently than is presently possible. An adjustable speaking rate, in conjunction with a computerized index or other method of content addressing could make reading almost as easy as the browsing we take for granted when we pick up a book (example 36 of the Appendix).

Another application is in the area of aids in the workplace. Of the 1.4 million visually impaired, many are elderly. However, 37 000 are children below 18, and 360 000 are between 18 and 64. Of these, about 106 000 are employed, according to the National Center for Health Statistics. Those blind individuals who work in an office environment could increase their productivity and become less dependent on a sighted co-worker if some sort of "talking text editor" computer system were available. Currently available systems are reviewed in *Aids and Appliances Review* (McGillivray, 1983).

D. Medical applications

Most medical applications are no different from other business applications, in that large health maintenance organizations employ centralized computer-based records on

patients that have to be accessed by phone when a doctor is not near a computer terminal. Attempts to use text-to-speech capabilities in novel ways have led to a computer system that tracks compliance in an experimental hypertension treatment program at Boston University Medical Center (Friedman, private communication). The computer calls each patient every day, and uses DECtalk to ask whether medication has been taken and whether any adverse side effects have occurred. The computer then calls a doctor if the patient's telephone keypad response indicates a problem.

Another potential application is an expert system for medical consultation between a doctor and a computerized data base. Those involved in artificial intelligence research have begun to amass large data bases on relations between symptoms and diseases. They hope ultimately to be able to reason logically, suggest additional tests, and deduce disease as well as the average practitioner—taking advantage of the superb memory capabilities of computers in order to consider rare clusters of symptoms that many doctors have not encountered in their practice. Text-to-speech telephone access could make such systems widely accessible and inexpensive.

VI. CONCLUSIONS

Text-to-speech conversion is a new technology with a rapidly changing set of capabilities and potential applications. The best of the current systems are quite intelligible, but suffer from a number of deficiencies that are often grouped under the catch-all term "lack of naturalness." In this article, we have identified many areas where rules and table values can be incrementally improved in the future to achieve more natural and more intelligible speech output from text-to-speech systems. As a consequence, these systems should become more acceptable to a wide range of users.

We have also identified several more basic problems that impede progress in certain areas of the text-to-speech conversion process (and also impact adversely on progress in other areas of speech science and technology). The first has to do with fitting spectral data obtained from female voices into the framework of current formant synthesizer models. For breathy vowels, the fit is not particularly good (recall Fig. 13), and it appears that some of the spectral deviations caused by tracheal coupling have perceptual importance.

It may be worthwhile to speculate on ways in which this problem might be resolved. Ideally, a new formant synthesizer model will be suggested that is slightly more complex, but still practical to implement. For example, an extra pole, or pole-zero pair might be made available to match extra spectral prominences that are observed. In this scenario, a way will be found to relate speech data from female voices to model parameters, so that a data collection effort will result in effective rules for controlling the new synthesizer model.

I suspect that the solution will not be that simple. If true, we may have to wait for speech science to provide better answers to some basic questions. The first point to note is that the acoustic theory of speech production, whether simplified or made complex by the introduction of better models of the larynx, trachea, and source–filter interactions, is not intended to be a model of the parameters directly controlled when we speak, nor of the parameters directly involved in the perceptual decoding of speech. The theory is a description of the acoustic behavior of a mechanical system. Therefore, efforts to relate observed spectral data from real female talkers to formant frequencies and other acoustic parameters of the theory have no *a priori* reason to succeed, and actually stand a good chance of failure, in part because there are too many model parameters compared with available spectral details (especially for talkers with high fundamental frequencies). Are we in a situation where it is possible to collect spectral data, yet be unable to relate it unambiguously to the underlying generation process, or to the processes of speech perception or articulation?

If this characterizes the present state of speech science, and I think it does, then the real bottleneck is the absence of a satisfactory perceptual theory to account for listeners' behavior in terms of observable spectral or waveform details. That we are far from such a theory is obvious, but how to go about attaining one is less clear. Attempts to mimic the steps believed to occur during the encoding stages of peripheral auditory processing are attractive as a first step, but it is unlikely that this encoding alone will be able to explain all of the fundamental perceptual skills that come naturally to humans, but not to speech recognition devices. Even the simplest of objectives, such as being able to categorize static critical-band spectra of vowels on the basis of a distance metric (Bladon and Lindblom, 1981), or to relate pairs of vowel spectra in terms of phonetic similarity (Klatt, 1982c), are well beyond our capabilities and understanding. Figure 34 shows pairs of critical-band spectra of vowels similar to/a/ that illustrate some of the difficulties encountered by a Euclidean metric. Spectral changes that affect peak locations are phonetically more important than other changes, even for low-pitched male voices synthesized to conform to the all-pole model of the vocal tract transfer function. But efforts to interpret critical-band spectra in terms of peak locations are thwarted in higher-pitched voices because individual harmonics are resolved, and breathy vowels introduce unexpected extra peaks. So long as we cannot always interpret spectral data from high-pitched voices in terms of formant parameters, or characterize the perceptual implications of spectral details, it is very likely that a synthetic female voice will remain an elusive goal, as may some aspects of the perceived naturalness of all male and female voices created by rules.

The second set of fundamental problems that we have identified arises when contemplating the creation of "natural" rule systems that manipulate articulatory structures. Where are the data that might facilitate creation of realistic models and model behavior? The acoustic consequences of any articulation depend on the cross-sectional area of the tube that is formed, and precision of specification is most important in locations of narrow constrictions. However, x-ray data, which are sparse, give only rough outlines in two dimensions, from which cross-sectional area must be inferred. And x-ray data do not characterize the masses and

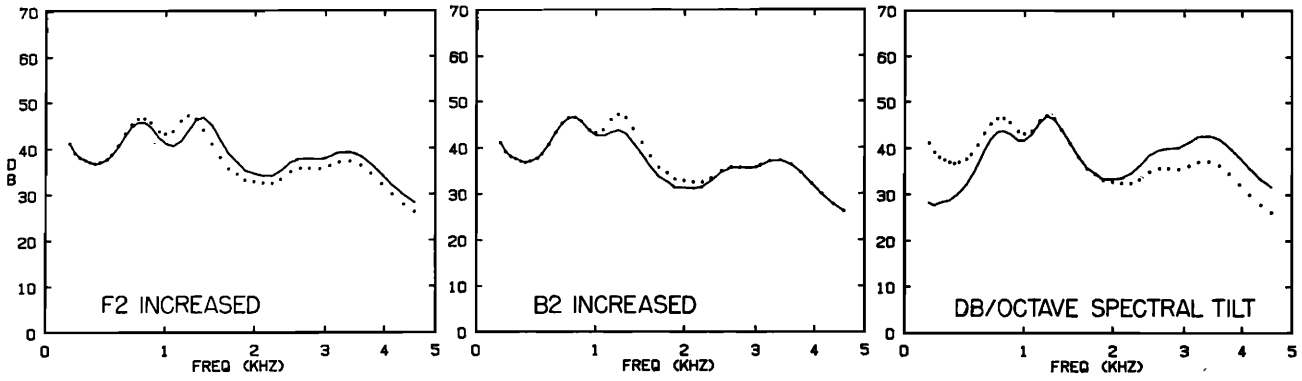

FIG. 34. Critical-band spectra of pairs of vowels that differ in terms of formant frequency location, formant bandwidth, or spectral tilt. Euclidean distance between solid and dotted curves does not reflect phonetic similarity between vowel pairs: An $F2$ increase creates a large change in judged phonetic difference, a B2 change is hard to hear at all, and a spectral tilt change is very audible, but does not affect judged phonetic similarity. Locations of energy concentrations seem to be of prime importance for phonetic categorization, but this hypothesis is difficult to maintain for high-pitched breathy vowels, see text.

degrees of freedom of the component articulators that make up the speech production system, nor the control constraints and strategies. There are a large number of researchers active in this area who make use of a wide range of devices to measure mechanical motions of individual articulators, and even EMG signals in individual muscles, but few if any of these scientists are pursuing a goal directly related to the assembly of synthesis rules for English syllables and sentences. Theorizing on the potential advantages of articulatory models in linguistics is another active area, but until such time as feature implementation rules become the central focus of effort (Goldstein and Browman, 1986), resulting in models detailed enough to reveal the immensity of the control problem, such theorizing is of marginal value to us. In summary, it seems that the study of basic processes of speech production and speech perception is crucial to progress, but is only in its infancy. Support of these activities will ultimately provide us with new insights and technical abilities. In the other direction, text-to-speech conversion is an excellent focus for sharpening the questions asked in basic research.

Let me quickly enumerate those areas where improvement in text-to-speech system performance is possible and reasonably straightforward. In all of the current systems, text analysis errors of many sorts are still possible. Deficiencies which have been identified in this article are summarized in Table XIII. The formatting routines may not be primed to deal with unusual letter or number strings. The word pronunciation routines have a certain probability of error in dealing with unfamiliar words, and this error rate tends to go up when dealing with foreign words and proper names. The syntax analysis routines may not be able to properly derive phrase structure for some sentences, or they may be unable to choose between two alternative pronunciations of an ambiguous orthographic word. These errors of text analysis are moderately frequent, occurring in as many as a third of the sentences of running text. Incremental improvements to formatting routines, augmentations to ever larger morpheme dictionaries (Coker, 1985), and additional parsing heuristics should lead to improved performance in this area. On the other hand, high performance syntactic analysis may turn out to require semantic knowledge, which would imply very large data structures and programs that may not be available for some time.

The problems remaining in the synthesis algorithms of text-to-speech systems are also listed in Table XIII. If one makes a spectrogram of a sentence produced by a text-to-speech system, and compares it with a sentence read by the person whose speech formed the basis for system development, it is easy to see ways in which the two acoustic patterns

TABLE XIII. Research issues for improving text-to-speech systems.

TEXT ANALYSIS
Text formatting
• programming what readers know about standard formats and abbreviations
Syntax/semantics
• syntactic analysis specifically for text-to-speech
• bootstrapping semantic information
Phoneme/stress prediction
• large-scale morphemic decomposition
• proper names
SYNTHESIS
Prosodics
• new systems of rules for f_0 control, improved duration rules
• mechanisms for getting variety into the rules
Phonology
• additional details concerning sentence-level phonetic recoding
Acoustic–phonetics
• segmental intelligibility
• detailed cues that may contribute to naturalness
Voice quality
• voicing source and tracheal coupling characterization for female voices
• source control as a function of time
APPLICATIONS
Technology transfer
• getting this technology into aids for the handicapped

differ. It is less easy to tell whether individual differences are perceptually important, but if one has some idea of discrimination limits, the perceptual salience of various speech cues, and the articulatory basis of acoustic discrepancies, then good guesses can be made as to the specific rules needed in the future. In this sense, all of the systems are amenable to incremental improvements so long as their designers have sufficient patience to follow this cookbook method of uncovering acoustic deficiencies.

Part of this process might even be automated. Holmes (1984) describes an effort to automatically time align a sentence with its synthetic imitation produced by rule, and then incrementally adjust formant frequency table values in the Holmes *et al.* (1964) rule program until natural and synthetic utterances are maximally similar. If the rules are correctly formulated and complete, such optimization procedures should result in improved imitations of other sentences as well. However, before such optimization efforts realize their full potential, many additional rules appear to be needed at the segmental level, e.g., to derive nuances of vowel quality change as a function of stress and phonetic environment. In the absence of a correct rule framework, automatic training will simply fail to converge, no matter how much data are supplied.

Text-to-speech programs and research may begin to have an influence on the way phonologists and phoneticians view phonetics and phonemic theory. These linguists have traditionally been reluctant to ascribe psychological reality to the phoneme, preferring to rely on distributional properties of observed sounds as a basis for theorizing (see Fry, 1974 for a good review). To the extent that speech generation programs begin to look like models of human behavior, their representations of language processes and units may become the cornerstones of new linguistic theories. If a synthesis-by-rule program can attract theoretical linguists to the problems inherent in specification of feature implementation rules, and thereby better couple their insights to the problem of allophonic variation, acoustic–phonetic detail, and timing of phonetic events, it is possible that real progress can be made in both engineering and linguistics. At the very least, it can be expected that these programs, in modifiable form, will become a part of the experimental facilities of modern phonetics laboratories, and will influence future generations of students in ways that are hard to predict.

In a similar vein, it is difficult to estimate the impact on the general public of computers that speak and listen. Talking machines may be just a passing fad, but the potential for new and powerful services is so great that this technology could have far reaching consequences, not only on the nature of normal information collection and transfer, but also on our attitudes toward the distinction between man and computer.

It is sometimes said that speech synthesis is not only *easier* than automatic speech recognition, but also that the field is so mature that the remaining problems are minor and scientifically uninteresting. I hope that this review has tended to dispel this view by pointing to specific areas where basic knowledge is lacking, and significant progress can still be made.

ACKNOWLEDGMENTS

Preparation of this review was supported in part by an NIH grant. I am very grateful to Ignatius Mattingly, John Holmes, Jared Bernstein, Osamu Fujimura, Stefanie Shattuck-Hufnagel, and David Pisoni for numerous suggestions based on an earlier draft.

APPENDIX: DEMONSTRATION

The enclosed 33 $\frac{1}{3}$-rpm recording contains illustrations of some of the milestones in the development of systems for text-to-speech conversion. For convenience in locating and listening to examples as they are described in the text, it may be desirable to transfer the recording onto a cassette tape. The assistance of H. David Maxey, Michael Hecker, John Holmes, Patrick Nye, Joe Olive, and James Flanagan in assembling these materials is gratefully acknowledged. My thanks also go to Kenneth Stevens, who served as narrator.

The record has been inserted inside the back cover of this issue.

Part A: Development of speech synthesizers

The objective of early research on speech synthesis was to test whether the synthesizer design is capable of high-quality imitations of human voices.

1. The VODER of Homer Dudley, 1939.

Dudley of AT&T Bell Laboratories designed a speech synthesizer known as the "Voder" (Dudley *et al.*, 1939). It was demonstrated at the 1939 World's Fair in New York.

2. The Pattern Playback designed by Franklin Cooper, 1951.

The Haskins Laboratories Pattern Playback (Cooper *et al.*, 1951) was designed to permit converting back into sound the patterns observed on broadband sound spectrograms.

3. PAT, the "Parametric Artificial Talker" of Walter Lawrence, 1953.

Lawrence (1953) of the Signals Research and Development Establishment, Christchurch, England, designed the "PAT" ("Parametric Artificial Talker") parallel formant synthesizer. It was first demonstrated at a conference in London in 1952.

4. The "OVE" cascade formant synthesizer of Gunnar Fant, 1953.

Fant (1953) of the Royal Institute of Technology in Stockholm, Sweden designed a cascade formant synthesizer ("OVE I"). It was demonstrated at the same London conference in 1952.

5. Copying a natural sentence using Walter Lawrence's PAT formant synthesizer, 1962.

Tony Anthony and Walter Lawrence attempted to match a natural recording using an updated version of PAT (Anthony and Lawrence, 1962). Demonstrated at the 1962 Stockholm Speech Communication Conference. Compare with the OVE II version of the same utterance, next.

6. Copying the same sentence using the second generation of Gunnar Fant's OVE cascade formant synthesizer, 1962.

Gunnar Fant attempted to match a natural recording using OVE II (Fant and Martony, 1962). Demonstrated at the 1962 Stockholm Speech Communication Conference. Compare with the PAT version of the same utterance, above.

7. Comparison of synthesis and a natural sentence, using OVE II, by John Holmes, 1961.

Holmes (1961) of the Joint Speech Unit of the British Post Office used the OVE II synthesizer to generate a close copy of a natural sentence.

8. Comparison of synthesis and a natural sentence, John Holmes using his parallel formant synthesizer, 1973.

Holmes did essentially the same thing in 1973, using a more complex parallel formant synthesizer of his own design (Holmes, 1973). Demonstrated at the 1972 IEEE Conference on Speech Communication and Processing, Boston.

9. Attempt to scale the DECtalk male voice to make it sound female.

The DECtalk "Perfect Paul" male voice has been modified by scaling f_0 by a factor of 1.7 (ap = 204, pr = 170), by scaling all formant frequencies by a factor of 0.85 (hs = 85) and removing the fifth formant (f5 = 2500, b5 = 2048), by increasing the open quotient of the glottal waveform using the "richness" variable (ri = 0), and by decreasing the output level slightly to avoid overloads (lo = 81). These manipulations are not sufficient to turn Paul into a convincing female speaker.

10. Comparison of synthesis and a natural sentence, female voice, Dennis Klatt, 1986b.

A synthetic copy of a female speaker producing (1) a sentence and (2) an utterance in which each syllable of "Steve eats candy cane" is replaced by [ʔa] is compared with the original recording (Klatt, 1986b).

11. The DAVO articulatory synthesizer developed by George Rosen at M.I.T., 1958.

The DAVO ("Dynamic Analog of the VOcal tract") circuit designed by Rosen (1958) at M.I.T., augmented by a nasal tract designed by Hecker (1962), was controlled by a tape recording of control signals created by hand by Kenneth Stevens and Arthur House. The demonstration occurred at the fall meeting of the Acoustical Society of America in 1961.

12. Sentences produced by an articulatory model, James Flanagan and Kenzo Ishizaka, 1976.

Flanagan and Ishizaka (1976) of the AT&T Bell Telephone Laboratories used an articulatory synthesizer to generate two sentences, using control data derived from the Coker *et al.* (1973) text-to-speech system. A two-mass model of the vocal cords was employed, and turbulence noise was injected automatically whenever the Reynolds number became large at the larynx, or at a constricted section of the vocal tract.

13. Linear-prediction analysis and resynthesis of speech at a low-bit rate in the Texas Instruments Speak-'n-Spell toy, Richard Wiggins, 1980.

Wiggins (1980) designed a low-cost linear-prediction synthesis chip to take advantage of the ability of linear prediction to represent critical spectral and temporal aspects of speech waveforms efficiently.

14. Comparison of synthesis and a natural recording, automatic analysis–resynthesis using multipulse linear prediction, Bishnu Atal, 1982.

Atal of the AT&T Bell Laboratories demonstrated a new formulation of linear prediction, known as multipulse LPC (Atal and Remde, 1982) at the 1982 Paris ICASSP.

Part B: Segmental synthesis by rule

The first synthesis-by-rule programs concentrated on the development of rules for phonemic synthesis, and did not include rules for the automatic specification of phoneme durations and fundamental frequency. Since prosody was specified by hand to match a natural recording, these demonstrations sound significantly better than they would if all information had been derived by rule.

15. Creation of a sentence from rules in the head of Pierre Delattre, using the Haskins Pattern Playback, 1959.

A stylized spectrogram of the desired sentence was painted on a transparant plastic plate by Pierre Delattre, and then played by the Haskins Pattern Playback.

16. Output from the first computer-based phonemic synthesis-by-rule program, created by John Kelly and Louis Gerstman, 1961.

Kelly and Gerstman (1961, 1962) of the AT&T Bell Laboratories demonstrated the first phonemic synthesis-by-rule program in 1961 at a meeting of the Acoustical Society of America.

17. Elegent rule program for British English by John Holmes, Ignatius Mattingly, and John Shearme, 1964.

Holmes *et al.* (1964) of the Joint Speech Research Unit in England demonstrated an impressive phonemic synthesis-by-rule program for British English at the fall meeting of the Acoustical Society of America in Ann Arbor, 1963.

18. Formant synthesis using diphone concatenation, by Rex Dixon and David Maxey, 1968.

Dixon and Maxey (1968) of IBM at Research Triangle Park demonstrated a diphone concatenation method for construction of control parameter time functions for a formant synthesizer at the 1967 M.I.T. Conference on Speech Communication and Processing.

19. Rules to control a low-dimensionality articulatory model, by Cecil Coker, 1968.

Coker (1968) of AT&T Bell Laboratories created a method of generating speech from an articulatory model. The system was demonstrated at the 1967 M.I.T. Conference on Speech Communication and Processing.

Part C: Synthesis by rule of segments and sentence prosody

The next synthesis-by-rule programs include a complete set of rules for going from phonemes, stress marks, and some syntactic information to an output speech waveform.

20. First prosodic synthesis by rule, by Ignatius Mattingly, 1968.

The synthesis-by-rule program of Mattingly (1966; 1968) of the Haskins Laboratories was demonstrated to accompany his Ph.D. thesis.

21. Sentence-level phonology incorporated in rules by Dennis Klatt, 1976.

Klatt (1976b) of the M.I.T. Speech Communication Group created a phonological component to generate segmental durations and a fundamental frequency contour, as well as sentence-level allophonic variation, from a phonemic input augmented with stress and syntactic symbols.

22. Concatenation of linear-prediction diphones, by Joe Olive, 1977.

Olive (1977) of AT&T Bell Laboratories controlled a linear-prediction synthesizer from stored reflection coefficients for a set of diphones. The system was demonstrated at ICASSP-77. The recording is from about 1980, and includes prosodic rules provided by Liberman and Pierrehumbert.

23. Concatenation of linear-prediction demisyllables, by Cathrine Browman, 1980.

A synthesis-by-rule program with prosodic rules, called Lingua, was designed by Browman (1980) of AT&T Bell Laboratories, using the demisyllable inventory collected by Fujimura and Lovins (1978). Demonstrated at ICASSP-80.

Part D: Fully automatic text-to-speech conversion

24. The first full text-to-speech system, done in Japan by Noriko Umeda *et al.*, 1968.

The first demonstrated text-to-speech system for English was created by Umeda *et al.* (1968) of the Electrotechnical Laboratory in Japan, and was based on an articulatory model. It included a syntactic analysis module with sophisticated heuristics. Demonstrated at the 6th International Congress on Acoustics, in Tokyo in 1968.

25. The first Bell Laboratories text-to-speech system, by Cecil Coker, Noriko Umeda, and Cathrine Browman, 1973.

Coker *et al.* (1973) of AT&T Bell Laboratories demonstrated a text-to-speech program based on the Coker (1967) articulatory model. The system was demonstrated at the 1972 International Conference of Speech Communication and Processing in Boston.

26. The Haskins Laboratories text-to-speech system, 1973.

The Haskins Laboratories text-to-speech system (Cooper *et al.*, 1973) used the Mattingly (1968) phoneme-to-speech rules coupled with a large dictionary.

27. The Kurzweil reading machine for the blind, Raymond Kurzweil, 1976.

Kurzweil (1976) began selling a reading machine with an optical scanner in the late 1970s. The system was demonstrated on the CBS evening news.

28. The inexpensive Votrax Type-n-Talk system, by Richard Gagnon, 1978.

The Votrax low-cost Type-n-Talk text-to-speech system combines a single-chip synthesis-by-rule program and formant synthesizer (Gagnon, 1978) with a version of the Elovitz *et al.* (1976) letter-to-sound rules. It was demonstrated at the 1978 ICASSP Conference.

29. The Echo low-cost diphone concatenation system, about 1982.

The Echo low-cost text-to-speech system concatenates linear-prediction diphones using the Texas Instrument's TMS-5220 linear prediction synthesizer chip.

30. The M.I.T. MITalk system, by Jonathan Allen, Sheri Hunnicutt, and Dennis Klatt, 1979.

The MITalk-79 laboratory text-to-speech system, developed at the Massachusetts Institute of Technology by Allen *et al.* (1979, 1987) and many others. The system was demonstrated in its final form at the 1979 meeting of the Acoustical Society of America in Boston.

31. The multi-language Infovox system, by Rolf Carlson, Bjorn Granström, and Sheri Hunnicutt, 1982.

The Infovox commercial text-to-speech system (Magnesson *et al.*, 1984) is an implementation of the Carlson *et al.* (1982a) multilanguage system that was developed at the Royal Institute of Technology in Stockholm by Rolf Carlson *et al.* Versions of the system were demonstrated in 1976 and 1982 at ICASSP conferences.

32. The Speech Plus Inc. "Prose-2000" commercial system, 1982.

The Prose-2000 commercial text-to-speech system was first developed in conjunction with a reading machine for the blind project at Telesensory Systems by James Bliss and his associates (Goldhor and Lund, 1983; Groner *et al.*, 1982). The recording is of Version 3.0 of the software.

33. The Klattalk system, by Dennis Klatt of M.I.T. which formed the basis for Digital Equipment Corporation's DECtalk commercial system, 1983.

The Klattalk (1982a) laboratory text-to-speech system software was licensed to Digital Equipment Corporation as a basis for the commercial DECtalk text-to-speech system announced in 1983. The recording is of Version 3.0 of the DECtalk software.

34. The AT&T Bell Laboratories text-to-speech system, 1985.

A new AT&T Bell Laboratories laboratory text-to-speech system (Olive and Liberman, 1985) uses the Olive (1977) diphone synthesis strategy in combination with a large morpheme dictionary (Coker, 1985) and letter-to-sound rules (Church, 1985). The laboratory system was demonstrated at a 1985 meeting of the Acoustical Society of America.

35. Several of the DECtalk voices.

Examples of some of the voices provided by the *DECtalk* text-to-speech system: (1) Beautiful Betty, (2) Huge Harry, (3) Kit the Kid, (4) Whispering Wendy.

36. DECtalk speaking at about 300 words/minute.

Example of using the DECtalk speaking rate command to skim material at a rapid rate. The nominal speaking rate has been set to 350 words/min, [:ra 350], although this 51-word passage took 11 s to speak, indicating an effective rate slightly under 300 words/min.

[1]For an application requiring a limited set of sentences with known struc-