TOEFL®

Monograph Series

MS - 13
SEPTEMBER 1999

A Review of Computer-Based Speech Technology for TOEFL 2000

Jill C. Burstein
Randy M. Kaplan
Susanne Rohen-Wolff
Daniel I. Zuckerman
Chi Lu

ETS Educational Testing Service

A Review of Computer-Based Speech Technology for TOEFL 2000

Jill C. Burstein, Randy M. Kaplan, Susanne Rohen-Wolff,
Daniel I. Zuckerman, Chi Lu

Educational Testing Service
Princeton, New Jersey
RM-99-5

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

To obtain more information about TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org

Web site: http://www.toefl.org

Foreword

The TOEFL® Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language program development efforts. As part of the foundation for the TOEFL 2000 project, a number of papers and reports were commissioned from experts within the fields of measurement and language teaching and testing. The resulting critical reviews and expert opinions have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project is a broad effort under which language testing at ETS will evolve into the 21st century. As a first step in the evolution of TOEFL language testing, the TOEFL program recently revised the Test of Spoken English (TSE®) and announced plans to introduce a TOEFL computer-based test (TOEFL CBT) in 1998. The revised TSE, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The TOEFL CBT will take advantage of the new forms of assessments and improved services made possible by computer-based testing while also moving the program toward its longer-range goals, which include

- the development of a conceptual framework that takes into account models of communicative competence
- a research agenda that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

It is expected that the TOEFL 2000 efforts will continue to produce a set of improved language tests that recognize the dynamic, evolutionary nature of assessment practices and that promote responsiveness to test user needs. As future papers and projects are completed, monographs will continue to be released to the public in this new TOEFL research publication series.

TOEFL Program Office
Educational Testing Service

Abstract

Computer-based speech technology, the capability of a computer system to accept and process spoken language, is considered a potentially super-enabling technology for computer users. Once a computer can adequately "understand" spoken language, the accessibility of computers increases by many orders of magnitude. As part of our on-going effort to examine enabling and important technologies, we have undertaken with this study to review the state of the art in computer-based speech technology in the context of the Test of English as a Foreign Language (TOEFL®) testing program. Our goal in this study is to assess the readiness of various computer-based speech technologies for this testing program. This paper focuses on desktop applications for speech recognition and speech synthesis.

Table of Contents

List of Tables

Introduction

The market for computer-based speech understanding systems currently comprises about 700 firms (Faulkner Technical Reports, 1994). It is one of the fastest growing sectors in telecommunications. The purpose of this report is to assess how commercially available computer-based speech understanding technology can be useful in the TOEFL 2000 program. In this investigation we are interested in ascertaining if: (a) the state-of-the-art computer-based speech recognition technology is sophisticated enough for applications to the TOEFL testing program, (b) speech-based technology could be used for service-related purposes within the TOEFL program, and (c) speech-based technology could be used to analyze oral stimuli.

Computer-based speech technologies consist of several different types of applications. The two major types of applications are *speech recognition* and *speech synthesis*. Speech recognition is the process by which a computer system "understands" spoken language. Speech synthesis occurs when a computer is simulating a human speaker. Both of these application types may have a place in the TOEFL testing program. Integrating computer-based speech technology into the testing process could enhance the testing of spoken English. After discussions with members of TOEFL program administration and test development, several ideas for speech-based testing applications arose. They include a computer-based testing system that records and analyzes verbal examinee responses to items or telephony-based procedures in which low stakes practice tests would be available to potential test takers over the phone. Speech synthesis could be used to deliver feedback to the examinee.

Service applications, such as course registration that use speech-based technology are already being implemented in a number of colleges and universities nationwide. Registration for TOEFL examinations could be expedited and managed using computer-based speech technology in a similar manner to potentially reduce the overhead incurred in the registration process.

This study investigated several commercially available speech-based technologies. The systems we evaluated are based on desktop computer technology. Due to the length of this study, this evaluation was conducted on a small, but representative sample of state-of the-art desktop systems. Systems were chosen because they were among the top-ranked personal computer-based, speech technology applications. The systems we evaluated were: *DragonDictate*® by Dragon Systems, Inc., *Kurzweil VOICE*™ for Windows by Kurzweil Applied Intelligence, Inc., and a text-to-speech synthesis system, *DECTalk*® *PC* 4.2, by Digital Equipment Corporation.

Computer-Based Speech Technologies

Brief Introduction

Speech technologies are based on a combination of acoustics, linguistics, and statistics theories and involve state-of-the-art analog and digital signal processing technology. Research in all disciplines that feed into speech technology has made headway in the last 30 years. The speech technologies discussed in this paper are speech recognition, speech synthesis, and interactive voice response.

Speech Recognition

As defined by Pelton (1993), speech recognition is the process of deriving either a textual transcription or some form of meaning from spoken input. In a speech recognition system, speech is collected using a microphone or telephone and transformed into digital data. This process is called *digitizing*. Once digitized, the data are analyzed and converted into text and/or a related meaning representation.

Speech recognition systems can be *speaker independent*, *speaker dependent*, or *speaker adaptive*. A *speaker independent* system can theoretically recognize the speech of any person. This type of system requires no training on the part of the user. Training is the process of familiarizing a system with the characteristics of an individual's voice. A *speaker dependent* system is trained on the voice of a single user. A *speaker adaptive* system requires training, but can be trained for multiple users. Depending on input constraints, such as vocabulary size and rate of speech, each category can be further subdivided into small and large vocabulary systems and isolated from continuous speech recognition systems. There is an inverse relationship between recognition performance-measured as word error rate–and both vocabulary size and speech rate (see Table 1).

Speech recognition systems intended to recognize isolated words from a limited vocabulary are quite reliable. Telephony-based voice response systems of this type are widely used by both the government and corporations. They might allow rotary phone users the same around-the-clock access to form ordering or information that touch-tone phone users enjoy. Some systems prompt for "yes" and "no" answers that replace pressing certain keys on the keypad, others also require the use of a limited vocabulary of digits and words (for example, the information systems of the Immigration and Naturalization Service, or New York State's Unemployment Insurance Division).

Systems trained to recognize a discrete set of vocabulary items that are not spoken in isolation (for example, credit card numbers, telephone numbers) are much more difficult to implement since the speech chunks that can be matched against the system's reference dictionary have to be established first. A speaker-independent system that recognizes a limited vocabulary (digits) of connected speech input is AT&T®'s Universal Card Customer Service. Information about balances and bills can be obtained by reading out loud the digits of one's credit card and giving other voice commands. Speech rate and volume are not constrained.

The telecommunications industry with its vast customer base has been a primary force in creating speaker-independent systems. Certain "custom" calls such as collect, person-to-person, calling card, and third-number calls can be handled without operator intervention by simply using the appropriate key word (collect, third-party, etc.) alone or in a sentence. The system is capable of keyword spotting and has so-called "barge-in" capability in

that it does not require the caller to wait for the end of a prompt. This system has been widely tested and is operational.

Systems trained to recognize the connected speech of any number of speakers and large vocabularies are still in their infancy. BBN Hark Systems Corporation speech recognition software has an active vocabulary of 2,000+ English words and supports continuous speaker-independent input. It is currently being tested by AT&T. IBM® is considering incorporating it into its IBM DirectTalk®/6000 voice response system (Voice News, August 1995). Carnegie Mellon's SPHINX system, a speaker-independent continuous speech recognition engine claims 97% recognition accuracy. It relies on a database of phonetic Hidden Markov Models that specify the probability of any speech sound in any given position, a 1,000 word dictionary, and a grammar of word pair transition probabilities.

The accuracy rates of current state-of-the art speech recognition systems are summarized in Table 1 (Rabiner, 1995).

Table 1
Recognition performance for several technologies and tasks

Technology	Vocabulary size	Task	Word error rate (%)
isolated words and phrases	10 digits 1,000 words	isolated digits basic English	.1 4.3
connected words	11 digits	connected digits	.2
continuous speech	1,000 words	database management	4.5
	2,000 words	airline reservations	2.3
	20,000 words	reading from the Wall Street Journal	11.3

Text-to-Speech Synthesis

Text-to-speech synthesis (TTS), (also referred to as *speech synthesis, speech production*, or *speech generation*) is the process of converting electronic text (generally ASCII text) into speech. TTS is well-suited for applications with large and rapidly changing databases. An example of current TTS technology is converting electronic mail messages into speech.

Interactive Voice Response (IVR)

Interactive voice response (IVR) is a telephony-based application that often combines several technologies, such as text-to-speech synthesis, speech recognition, and keypad (i.e. touch-tone) recognition. Telephony-based speech systems use the telephone as the means of capturing voice input. This is different from personal computer-based systems that typically use an external microphone to capture voice input.

IVR systems automate many of the telephone functions traditionally handled by a customer service representative. When callers use an IVR system, they are usually given a menu of options, and asked to use the touch-tones on the phone to select options, or to speak a particular word or number to trigger an option.

For example, when calling an insurance company, you might get a menu like:

> *"Press 1 for car accident claims."*
> *"Press 2 for home owners claims."*
> *"Press 3 for disaster claims."*
> *"Press 4 to speak to a claims specialist."*

If a speech recognition system is built into the IVR, an alternative menu like the following might be given:

> *"Say* ***yes*** *immediately after you hear the option you want:*
> *"accident claims..."*
> *"homeowners claims..."*
> *"disaster claims..."*
> *"I want to speak to a claims specialist."*

A Review of Computer-Based Speech Technology

Systems in use at ETS

During our investigation of computer-based speech technology, we learned that two telephony-based systems are currently in use at Educational Testing Service (ETS). Although these systems are not using any speech recognition capabilities, it is helpful to understand how these systems are being used, and what potential they may have for future applications. These systems are described in the next section.

Telephony-based interactive voice response applications

ETS uses the Periphonics Voice Response System for Scholastic Assessment Test (SAT®) and Graduate Record Examinations (GRE®) exam registration. The Periphonic system includes a speech recognition component that is not used in the present registration application. The telephone keypad is currently used for registration. The Periphonics Voice Response System for SAT and GRE exam registration is specifically designed to allow access only to domestic callers. Previous experience has shown that callers from other countries are often unfamiliar with IVR use, and end up making costly mistakes.

We mention the Periphonics Voice Response System here because it has a speech recognition component and ETS has already invested in this technology. Although we have not evaluated the system in this study, the existence of this system as part of ETS's infrastructure makes it a good candidate for further investigation.

The Network Services Division of ETS is currently working on a project for the United States Immigration and Naturalization Service (INS) called *The New Citizens Program.* For this project, the Network Services Division will be implementing an IVR with speech recognition capability for the purpose of deploying a test of spoken English for citizenship applicants. The exam will be administered by the field service offices. The system will include a speaker verification component to check that the caller is a valid examinee. This will be accomplished by comparing a recording of the examinee speaking his or her name to the name spoken at the time of the exam. Another speaker verification component will verify proctors. The INS will be able to call a central location to retrieve an examinee's test results using an IVR that requests the examinee's alien identification number. The system will include a test of spoken English that can be taken over the phone; a speech recognition system will be implemented to handle examinee responses.

Overall outlook for speech synthesis and speech recognition technology

Speech synthesis research is directed toward improving intelligibility and making the speech produced sound natural. The orthography of the input text generally does not supply the type of information that is important for natural-sounding speech, such as duration of a sound, emphasis on a syllable, lack of emphasis on another, and the intonation of a phrase or sentence.

The integration of telephone and speech processing technologies, and its potential for cost-effective retrieval and dissemination of information, is being actively pursued by the large telephone companies and marketed to banks and government agencies. Among the applications of speech recognition currently being investigated is speaker verification by a voiceprint, which is analogous to a fingerprint. This could be useful as a security measure in telephone banking or for other long distance transactions via telephone lines, or to augment existing security measures.

Speech recognition research into language identification, that is into systems that ascertain the language of a speaker via certain keywords and statistics, such as the occurrence of various sounds and sound combinations, could prove useful for international calls.

AT&T has investigated machine translation for limited domains with its VEST English/Spanish translator for banking and currency transfer. It uses a very limited vocabulary and operates in near-real time.

Integration of computer software applications, speech recognition, and synthesis technologies is also on the rise. Speech technologies are part of Application Programmer's Interfaces (APIs) and allow PC or Macintosh users to control the computer via speech as opposed to typed or double-clicked commands. Some of these technologies are bundled with speech interfaces for popular software applications. Speech recognition and synthesis tool kits for languages such as Visual Basic® allow software developers or system integrators to incorporate speech processing into new applications. English language instruction software products (*Learn to Speak English*, Lernout & Hauspie) provide an interactive learning environment with the capability to record and score the user's pronunciation. Educational word processing programs for children use speech synthesis to read back anything that is typed in.

The greatest challenge for speech recognition technology is to create a speaker-independent system with a large vocabulary that utilizes syntactically and semantically unconstrained input. Speech recognition systems rely on matching speech chunks against an internal reference inventory of such speech chunks. With an English vocabulary of well over 500,000 lexical items, matching speech chunks of word size against a model vocabulary is a sizable task, even if only a single speaker is involved. It becomes an unwieldy proposition when the idiosyncratic speech patterns and voice characteristics of a large number of speakers have to be considered. Therefore, instead of word-size segments, smaller units, such as syllables or sequences of two (dyads) or three (triads) consecutive phonemes, are compared with a model inventory of such units before being pieced back together into words. Establishing linguistically relevant subword units, and stringing them together to form larger units presupposes extensive linguistic analysis and a complete inventory of possible sequences of such units. The potential sentences that can be created in English do not form a finite set, hence speech recognition technology has to rely on means other than finite model inventories.

Statistically-based techniques are being further explored as a way to harness the problems encountered by speech recognition systems that deal with large unrestricted corpora and large user populations. They are helpful for estimating whether a speech chunk uttered by a given user represents, or matches, a speech chunk from the model inventory. These methods are also helpful for estimating the concurrence probabilities of two or more consecutive speech units. For instance, can the speech sound [n] follow [k]–the answer is yes, witness *I like Nancy*–and if so, under what circumstances? The answer is that the sequence is only permissible across syllable

or word boundaries, compare *acknowledge* and *knife*. Statistical methods aid in deciding whether "goes" or "hoes" is a likely successor to the article "the."

Electroacoustic research has been making headway. Most current speech recognition systems are telephony-based and make use of the rather narrow telephone bandwidths available. But experimentation with different bandwidths, and the use of finer-tuned microphones in conjunction with normalization techniques, should improve the performance of speech recognition systems by reducing the ambient noise and machine noise that impair signal quality. Low-cost digital speech processing technology for simultaneous compression and decompression of speech is now available and is included on microprocessors for PCs and even home appliances.

Implemented systems, though quite sophisticated, are still lacking when compared with the best speech recognition system of all, the human brain. And despite all research efforts, there are no systems as yet that can be said to truly "understand" spoken language like a human being does. Humans hear a connected strings of sounds, break it up into its component sounds, separate chunks of sounds into words, associate a syntactic function and a word sense with them, and compute the meaning of the string in a matter of milliseconds, as they bring other cognitive faculties into play. A goal of current research is to go beyond deciphering the speech signal and extracting some rudimentary information, toward mimicking human speech recognition. Future systems may well "understand" spoken language like a human being.

Personal Computer-Based Speech Technology for this Study

Two speech recognition systems and one speech synthesis system were selected for this evaluation. These systems were selected based on the vendor's reputation in this area of technology and their market presence with these products. Dragon Systems, Inc.'s *DragonDictate* and *Kurzweil VOICE for Windows* are competing speaker-independent PC-based speech recognition applications that are designed for computer-based dictation and Windows navigation. These systems allow a user to enter text by speaking into a microphone connected to a PC, and to use voice commands to control computer applications like *Microsoft® Windows.*

Different versions of *DragonDictate* are available with different-size recognition vocabularies. We chose a version with a 30,000-word vocabulary. A larger dictionary (60,000 words) is available, but the entries in this dictionary are highly specialized. *DragonDictate* can be purchased for $1,020.

Like *DragonDictate, Kurzweil VOICE for Windows* can be purchased with a 30,000 or 60,000-word vocabulary. In the case of *VOICE for Windows*, the additional 30,000 words are not domain-specific and add to the product's recognition abilities. We chose the larger vocabulary for our evaluation. *VOICE for Windows* can be purchased for $995.

Speech recognition systems are currently used as a tool by people with physical handicaps or injuries, and also as a tool for executives or professionals with limited typing skills. Using these technologies, we can envision two potential applications for testing speech, based on our discussions with TOEFL program direction and test development. These are: (a) *short-answer, low stakes, over-the-phone, practice tests*, and (b) *short-answer speaking tasks that tested for response content.* We will evaluate *DragonDictate* and *VOICE for Windows* to determine their capability to recognize the speech of the TOEFL population.

The speech synthesis system, *DECTalk PC*, Version 4.2 developed by Digital Equipment Corporation is a state-of-the-art, text-to-speech synthesis software for the PC. *DECTalk* converts ASCII text into speech. *DECTalk* can produce nine predefined voices (four males, four females, and one child). It can speak at rates of 75 to 650 words per minute. The cost of this system is $1,690.

DECTalk has proven useful for a variety of applications. It is used by individuals with reading or speech impairments. *DECTalk* has been incorporated into system management software to generate spoken alerts that warn users of potential problems and also to allow users to screen and administer electronic mail messages (e-mail) more quickly.

Text-to-speech systems might be used to relay feedback to examinees taking practice tests over the phone. For instance, the text-to-speech system could use text from the examinee's response to relay feedback, such as "*Your answer X is incorrect, please try again.*" It might also be used for presenting stimuli to examinees during a testing session. In order for text-to-speech systems to be used in this way, it is important that the speech they produce be intelligible to a wide range of individuals. In this study we will examine the intelligibility of this text-to-speech system.

Evaluation of Systems

Speech recognition technology must satisfy at least two basic requirements if it is to be used in practical real-world applications. First, it must convert speech to text accurately. Second, it must be speaker-independent so that it can accommodate the unique speech characteristics of each user.

We chose to measure accuracy by establishing the ratio of correctly recognized text words to the total number of words in a text. We also chose a small, but diverse user population so that we could assess the system's performance with regard to male and female voices as well as to native and nonnative speakers. To compare the performance of *Kurzweil VOICE for Windows* with that of *DragonDictate,* we used the same text. Also, we used the same subject pool whenever possible.[1]

Kurzweil VOICE for Windows

Kurzweil VOICE for Windows promises recognition of at least 85% accuracy if it is used without any speaker training. Training the system to boost accuracy is highly recommended. This training, called "Enrollment," requires each user to record a sample vocabulary. The Enrollment window consecutively displays 400 word prompts that the user has to dictate to the system. The acoustic parameters of this sample are calculated and make up the unique user profile. The calculation phase takes about 90 minutes. This clearly suggests two test runs, one for each speaker before "Enrollment" (i.e., training) and another after training has established the subject's voice profile.

Data collection. Each subject was asked to read a short practice passage three times to get accustomed to the recognition speed of the system. Subjects were then instructed to read the main text twice before they had any training on the system. The speed at which *Kurzweil VOICE for Windows* recognizes spoken language is considerably slower than normal continuous speech. Speakers have to pause briefly between words while they read. Subjects were instructed to repeat each word of the passage until the word was recognized by the system. However, if recognition was not achieved by the fifth repetition of a word, subjects were asked to type in the word and enclose it in square brackets to indicate "no recognition."

To determine the impact of training on the system's effectiveness, subjects went through "Enrollment." The voice profile for each speaker was calculated based on the training vocabulary presented during Enrollment. Subjects were then asked to re-read both the practice passage and the main passage so that pre-training results could be compared to post-training results.

Text used for the evaluation. The main passage is given below. We chose a non-technical text about bats excerpted from *National Geographic.* The text had a good mix of monosyllabic and polysyllabic Germanic and Romance words.

---

[1] Scheduling and availability of individuals precluded our having the same set of individuals for both experiments with the speech recognition applications (see *Data collection, Text used for the evaluation, and Test population*).

"No stealth aircraft could be more sophisticated than the California leaf-nosed bat. It swoops so quietly through the desert night that it is called a whispering bat. Its eyes can spot a sleeping insect and its huge ears can pick up the sounds of a caterpillar's munching jaws. Only on the darkest of nights does this bat activate its ultimate detector. Through its nose it emits high frequency, low intensity echolocation signals created by contracting muscles in its larynx. Sound waves return to its ears after bouncing off doomed prey. This amazing bat is one of 44 North American species studied by the author. He has long emphasized the beneficial nature of bats, which feed voraciously on insect pests that yearly cost farmers and foresters billions of dollars in losses. Bats also pollinate plants and disperse their seeds. Although many myths have been dispelled, bats still need protection from vandals and from the growing practice of sealing up caves and mines that animals need to survive."

Test population. Table 2 summarizes the test population used for the evaluation of *Kurzweil's VOICE for Windows.*

Table 2
Test population for *Kurzweil VOICE for Windows*

	native	nonnative	total
male	2	2	4
female	3	3	6
total	5	5	10

All of the nonnative speakers had learned English in school and used English in their professional lives. Two used their native languages (Chinese, Spanish) as the primary language at home. The native languages of two of the other subjects were Spanish and German. The third subject was a native trilingual, whose languages were Chinese, English, and Malay. Table 2 summarizes the language characteristics of each subject used in the evaluation of *Kurzweil's VOICE for Windows.*

Table 3
Nonnative subjects gender and years of English instruction

Gender	Years of English instruction	Use of English at home
male i	10+	no
male m	10	yes
female c	13	no
female s	14	yes
female r	12	yes

Results. All fully recognized words, including homophones, were counted as correct. Specifically, the system was given credit when it identified *its* instead of *it's*, or *sealing* instead of *ceiling*, or *knows* instead of *nose*, *two* or *too* instead of *to*, and *one* or *won*.[2] Words that were only partially recognized, e.g. the singular *signal* instead of *signals, disperse* instead of *dispersed,*[3] or not recognized at all, were counted as unrecognized.

A number of text words were not in the Kurzweil dictionary, hence they could not be recognized correctly under any circumstances. Customizing the system dictionary for a given text by adding text words or inflected forms of text words is possible, but we chose not to do so for this study due to time constraints. Sometimes, the Kurzweil speech recognizer achieved only partial word matches. *Echolocation,* for instance, was rendered as *notation or accreditation*, with the correct suffix picked up. Once *echolocation* appeared as *relocation,* with the base word *location* correctly recognized. *Munching, foresters* and *voraciously,* were approximated by *launching/luncheon, oysters/fosters,* and *autiously/previously/originally,* respectively. Occasionally, these words were not recognized at all and had to be entered on the keyboard.

Three sample outputs are shown in Table 4. The correctly recognized words are in italics. The first two outputs represent the system used without training, the third, labeled Subject c, is Subject c's reading after training.

[2] It is clear that homophone recognition is not sufficient for real-world applications. More work needs to be done with regard to distinguishing appropriate word distribution either by syntactic or semantic means.

[3] *Signal* and *signals* differ only in one element, [s], but so do the words *no* and *low*. The difference in the former is grammatical, the latter lexical. The Kurzweil system does not analyze the input past the word level, hence it makes no sense to assign partial credit based on whether a misrecognized subunit of the word changes the grammatical function or the sense of the word.

Table 4
Text recognition samples for *Kurzweil VOICE for Windows*

Subject	Recognized Text
d	*no* stone *aircraft* good the *more sophisticated than this California* leave those that. *It* swaps *so quietly* referral *the desert night that it is called a)* that. *It's* highs camp *spot a sleeping insect, and its* inch *years can pick up* lifestyles *of a counselors* launching draws. *Only* one *the* artist *of nights does this* that *activate its detector.* Referral *its nose* diplomats I *frequency, low intensity* prohibition *signals* creative *by contracting muscles in its* marks. some ways *return to its ears after bouncing off doomed* credit. *This amazing* that *is one of 44 north American species studied by the* other. 8 *have long emphasized the beneficial nature of* baths, *which* he originally *on insect pests that nearly cost* workers *and* borders *millions of dollars in losses.* That's *also* following *plants and* dispersed *their* seats. *Although many myths have been* dispel, that's *still need protection from* pebbles *and from the growing practice of ceiling up* tapes *and* minds *that animals* Mead *to survive.*
c	*no* youth [aircraft] put the mall *sophisticated* they these *California* beef most that 8 soaps *so* decay no *the* assertion not that 8 *is* for the mystery that. 8 ice day *spot a* speaking infected, bank 8 *huge* years gain 8 often met *sounds* of a hectares monkey Johns. Owning home *the* Pakistan *of nights* thus these that *activate* 8 automate *detector.* So 8 *knows* 8 addicts *high frequency,* no integrity *signals* create *by* contacting *muscles* in 8 UNIX. Some beliefs region *to* 8 years *after bouncing off* doomed 3. The *amazing* that these *1 of* lethal not *American* visas stymied *by the author.* The *has long emphasized the beneficial* major *off* that, *which* fact [voraciously] *on* exact path *that* any *cost,* N fosters unions *of* daughters being bosses. That *also* eliminate inmates bank destroys *their* speed *although* mini these *have been* [dispelled], that steel the attaching *from* maintenance *and from the growing practice of* seeming *up* games *and* my *that animals* the *two survive.*
r	*No* skiers *aircraft* put the mall *sophisticated* they these *California* leave most that. *It* soaps *so quietly through the desert night that it is* cord *a* restrain that. *its eyes* 10 sport *a sleeping insect, and* 8 *huge* years cane take *up* lure some *of a* educators mounting jars. Army *on the* largest *of nights* thus these at aggravate 8 automate *detector. Through* 8 *knows it* ended *high frequency,* no *intensity* application *signals* create *by* contacting *muscles in its* merits. Some with *return to its* years *after bouncing of* pumped clay. These *amazing* that *is one of 44* not *America* dishes stymied *by the author. He has long emphasized the beneficial* major *of* that, *which* feet additionally *on insect pests that* nearly caused from *and* forests being of daughters in bosses. As *also* eliminate blamed *and* dispersed *their* speed. *Although many* knees *have* the disappeared, that steel meat production *from* lenders *and from the growing practice of* seeing *up* tapes *and* mine *that* animals meat *to survive.*

The correctly recognized words for each of the readings of the main passage never reached the 85% accuracy level claimed in the documentation, but the system reached 80% accuracy for a few

speakers. We took the average of the accuracy levels for the two readings of the main text before training and after training and recorded the change. We also calculated average accuracy levels for each group of speakers. The levels of accuracy attained with *Kurzweil VOICE for Windows* are shown in Table 5.

Table 5
Accuracy levels for native and nonnative male and female speakers using *Kurzweil VOICE for Windows*

native				nonnative			
	training		change		training		change
	untrained	trained			untrained	trained	
male j	74.35	78.18	3.825	male m	73.48	75.69	2.21
male d	72.10	78.73	6.63	male i	55.81	66.30	10.49
average	73.22	78.46	5.228		64.65	71	6.35
fem j	61.88	61.325	(.555)	fem c	41.72	57.46	15.73
fem e	58.84	67.125	8.28	fem r	55.80	71.27	15.47
fem l	77.35	79	1.66	fem s	54.69	69.34	14.65
average	66.02	69.15	3.31		50.74	66.02	15.28

The sample is too small to generalize reliably. We did note some trends, however. Male voices seem to be recognized better than female voices. According to Kurzweil, different algorithms are used to compute male and female voices that may contribute to recognition rate over gender. Native voices are recognized more accurately than nonnative voices, although the system does very well with nonnative male voices. Training improves recognition in the following scenarios: (a) improvement is slight for native males and females, and (b) improvement is quite pronounced for nonnative female voices. In this study, we have not researched why improvement occurs more with one group than another. Perhaps this could be done in a later study. Recognition accuracy for nonnative speaker female c in both untrained and trained trials are relatively lower than the accuracy for the other subjects. Our informal explanation for this is that female c has more difficulty with her pronunciation than the other speakers and that perhaps this is related by the system's ability to recognize her speech.

Critique of Kurzweil VOICE for Windows. Kurzweil VOICE for Windows is quite accurate when it comes to vowel recognition. Note in the recognition samples shown in Table 4 that many misrecognized words that share one or more of the vowel phonemes with the input words. A small sample is given on the next page, with the input words in italics and the output words next to them.

Input	Output as recognized by *VOICE for Windows*
could	good
be	the
bat	that
ears	years
prey	gray
losses	bosses
ultimate	alternate
sophisticated	investigated

Kurzweil VOICE for Windows is also quite accurate with regard to the recognition of the syllable structure and the syllable nuclei. Input and output words generally have matching numbers of syllables: *caterpillar* (4) and *educator* (4), or *studied* (2) and *stymied* (2). The vowels that form the nuclei of the syllable are often, though not always, correctly recognized: *draws (jaws), investigated (sophisticated)*. The Kurzweil speech recognizer is not as accurate with regard to consonant recognition. Voiced bilabial stops, such as [b] and [m] in *billions* versus *millions;* and labiodental fricatives [f] and [v]; sibilants [s]; and [z] in word final position with vital grammatical (plural markers, personal ending) functions, were frequently not recognized or misrecognized. Polysyllabic words, e.g. *sophisticated, detector, California*, with more vowel and consonant clues, were recognized more reliably than monosyllabic words.

Recognition improves with training. We had one nonnative subject train the system on each unrecognized or misrecognized word and subsequently read the passage one more time. The recognized text is shown in Table 6.

Table 6
Text recognition samples for *Kurzweil VOICE for Windows*
with specialized training

Subject	Recognized Text
c	*no* steel *aircraft could be more sophisticated* they *this California leaf nosed* that. *It swoops so quietly through the desert night that it is* cord *a whispering* that. *It eyes* gain *spot* her *sleeping insect, and* it *huge ears* 108 *up* their some *of a caterpillar's munching jaws. Only on the darkest of* 9 *does this* that *activate* it *ultimate detector. Through* it *knows it* inmates *high frequency,* now *intensity echolocation signals* create *by* contacting *muscles in* it very. *Sound* with *return to* it years *after bouncing off* home clay. *This amazing* that the *1 of 44* not *American* aegis started *by the author.* The *has long emphasized the beneficial* major *of* debt, *which* feet *voraciously on insect* has *that* DNA cross bombers *and foresters* is *of* daughters being *losses.* That *also* automate bland *and disperse their seeds. Although* naming live *have been dispelled,* that steel *need protection from* vendors *and from the growing practice of* viewing upper tapes *and mines that* edible *survive.*

Comparing Tables 4 and 6, which show Subject c's readings after quick "Enrollment" training and after thorough word-for-word training, respectively, one notices a dramatic increase in the number of correctly recognized words. The recognition rate jumped from 41.72% accuracy in an untrained state, to 57.46% after quick training, to 70.16% after thorough word-for-word training.

The system does not seem to make use of even the roughest syntactic or semantic analysis, but merely concatenates recognized words regardless of their lexical category. This results in output that is often nonsensical as illustrated in the sample outputs in Table 4 or Table 6.

Intelligibility. For a system to recognize 80% or even 85% of the text accurately is quite impressive. But is that enough for text intelligibility? We presented three *Kurzweil VOICE for Windows* outputs of varying degrees of recognition accuracy to a few colleagues who had not participated in testing the system and therefore did not know the text passage. We asked them to read the passage and briefly summarize what it was about. While the text with 41% correct vocabulary recognition was dismissed as unintelligible, and while both the 75% and the 80% outputs were perceived as markedly better, the readers had only an inkling that the text was about some animals, but could not say or even guess which ones.

DragonDictate for Windows

Data collection. DragonDictate strongly recommends that a user run "Quick Training," in addition to the preliminary voice profile, to enhance performance. We tested the system in an untrained state with one speaker. Recognition was so low that we chose not to include tests of the system in an untrained state. Instead, we followed *DragonDictate*'s suggestion and drew our conclusions from a trained set of speech recognition data. "Quick Training" consists of four successive training vocabularies, both general English vocabularies and command vocabularies that permit hands-free operation of word processing and spreadsheet programs, for a total of 745 words. The user is prompted for each word and may have to repeat a word a number of times before the system moves on to the next word. One may not have to train for all words in each of the training vocabularies; the total number of words in each vocabulary is reduced as the system adapts to the speaker. "Quick Training" takes approximately 30 minutes.

After completing "Quick Training," our users were asked to read a short text twice in order to get used to the recognition speed of *DragonDictate,* then to read twice the same longer text passage that we had used in the *Kurzweil VOICE for Windows* evaluation (refer to section Text Used for the Evaluation). As before, words had to be repeated until they were recognized by the system, but after the fifth repetition of a word, speakers were asked to type in the word and enclose it in square brackets to indicate that it had not been recognized. In addition to text words, punctuation marks ("comma," "period," "colon") and one formatting command ("new line") had to be dictated into the system. We used the same non-technical text about bats that we used for the *Kurzweil VOICE for Windows* evaluation.

Test population. The test population used for the evaluation of *DragonDictate* was almost identical to the one used for *Kurzweil VOICE for Windows.* We had one additional nonnative female speaker, and substituted

one native male speaker since one of the original male subjects was unavailable. The composition of the test population is summarized in Table 7.

Table 7
Summary of test population for DragonDictate

	native	nonnative	total
male	2	2	4
female	3	4	7
total	5	6	11

All the nonnative speakers had learned English in high school or college and used English in their professional lives. Three subjects used their native languages (Chinese, Spanish) as a primary language at home. Table 8 identifies the individual nonnative speakers by gender, by years of formal schooling in English, and by whether English is spoken at home or not.

Table 8
Years of English instruction and use of English at home

Gender	Years of English instruction	Use of English at home
male i	10+	no
male m	10	yes
female c	13	no
female w	8	no
female s	14	yes
female r	12	yes

Results. As before, we counted homophones, such as *knight* for *night*, and *knows* for *nose*, as correctly recognized words. The count also included punctuation marks since these were dictated as words. The accuracy rate is the percentage of correctly recognized words. For each subject we calculated the average accuracy rate from the two readings. Table 9 shows the average percentage of recognized words for each subject.

Table 9
Accuracy levels for native and nonnative male and female speakers

native		nonnative	
	word recognition (%)		word recognition (%)
male b	45.86	male m	43.93
male d	47.24	male i	44.48
average	46.55		44.21
fem j	41.72	fem c	30.11
fem e	33.98	fem w	18.79
fem l	56.36	fem s	40.33
		fem r	43.37
average	44.02		33.15

Average recognition rates for native and nonnative males and for native females are very similar. They are somewhat lower for nonnative females. Again, as with the Kurzweil rates, there seems to be a correlation between the recognition rate and a "heavy" accent. A larger sample would allow us to draw more precise conclusions.

Sample outputs of DragonDictate for Windows. Some sample recognition outputs for *DragonDictate* are shown in Table 10.

Table 10
Text recognition samples for DragonDictate

Subject	Recognized text
1	*no* elf *aircraft could be more sophisticated than this California* Wheat dues that. *It* so *quietly* grew *the* better *night that it is called a whispering* that. *It's* I had spots *a* leaking insects, *and its* improving within years had *pick up the* town *of a* actuary monthly cost. *Only on the* preface *of* night *does this* act *activate* it openness *detector.* grew it dues it even I reflecting, *low intensity* equities *signals* pp. I *contracting muscles* and it would. Found made turn *to its* years *after bouncing* cost with they. This BBC at *is* on of foreseeable port comparison *species* came *by the author. He has long* excise *the* benefit major *of* that, *which* he morbidity *on* insects S. *that yearly* cops *farmers and* orators *billions of dollars in losses.* That's *also* collects plant *and* first there seemed. Also *many* knits *have been* distilled, that ill *need* perfection *from* medical *and from the growing practice of* feeling up a *and* mind *that* angles *need to* buy.
d	non *Stealth aircraft* the *more sophisticated* and *this California* 89 F. In slips *so* lightly *through the* better than them in use all day historian at. *Its eyes* and stopped in *sleeping* insert, *and its* use years and in a *the of* any counselors in jobs. *Only* one *the* first *of nights does this* neck activity mix alternate *detector. Through its* notice in units by requesting, though density correlation *signals created by muscles in its* plants. So leaves referred *to its* years *after bouncing* a human right. *This amazing* and use *one of* 30 fourth *North* Arabic study *by the* other. *He has* one *& the beneficial* major *of* S., *which* the originally won second tests *that yearly* plus partners *and* orifice millions *of dollars in losses.* S. *also* fully *plants and* discourse *their* seats *although* the minutes *have been* excelled, S. *still* the action *from* animals *and from the growing practice of ceiling* a case *and* its neck *animals* the *two survive.*
r	*no* help as good the *more* dated men *this California* needs those that. in *so* slow Miami *through* the benefit *nights that its is called a* (that. *It's eyes* and *up a* peaking insects, *and its huge* years and in *up* the town of a factor questioning of. *Only on the* doctrine *of nights* as *this* act estimate *its ultimate* factor. *Through its nose* its heaven thy difference 8, load infancy" e3 I. Trenton muscle *in its* Warrens. Town weight turned *to its* years *after* pounding *of* doing gray. *This amazing* that *is one of 44 North* Meredith E. Penney I. Beale. *He has long* as 5% *nature of* that, reach he forged *on insects* that that's *yearly cost* commerce enforcement buildings *of dollars* England. At *also* quality when *and* first down the period *although many* this has *been* felt, at Bill needs action *from* medical *and from the growing practice of* dealing *up* a *and* my met" the *two* sides

The first of the sample outputs above (Subject 1) is substantially above the average accuracy rate. Samples d and r show average recognition, which hovers around 45% for native males and females and nonnative males.

Critique of *DragonDictate*. *DragonDictate* performs well with lexically closed, i.e. finite, classes of lexical items, such as prepositions, pronouns, conjunctions, and determiners. Content words, such as nouns, verbs, and adjectives, make up the open lexical classes and are not recognized as accurately.

DragonDictate concatenates words and names without calculating the syntactic or semantic probability of that word. It is less sensitive to the number and the phonetic makeup of syllables than the *Kurzweil VOICE for Windows*, which frequently "recognized" words that agreed in syllabicity and vowel structure with input words. Inversions of the type *spot* and *stop* showed again that the recognition of stops [t,p,k] is extremely difficult, as is the differentiation among fricatives [s,z,θ].

DragonDictate responds to training on individual words with considerably improved recognition accuracy. *DragonDictate* presents a "choice list" of words, or list of alternatives for the word just spoken. A subject can immediately correct the system by choosing the correct alternative from the "choice list" without reading it in again. If no alternative is presented, that is if the system has not been able to find a candidate word, subjects can type or dictate the spelling of the word to the system.

One speaker trained on the short test passage and recognition was near perfect after the third pass. The immediate feedback on the screen caused subjects to pronounce training words much more carefully and served as an informal pronunciation scoring device.

Intelligibility. *DragonDictate* after "Quick Training" ranks below the trained *Kurzweil VOICE for Windows* output with respect to accuracy rate, with an overall (native and nonnative, male and female) average accuracy rate around 42% as opposed to an overall average of 70% for *Kurzweil Voice for Windows*. It also ranks below the Kurzweil output with respect to intelligibility. Even the best output (see Table 10) provides listeners with no clue that the text is about bats.

DECTalk PC

DECTalk PC produces speech output from text input. We tested this system for the clarity and intelligibility of its output and for naturalness, i.e. whether or not the voice quality closely resembled a human voice.

Data collection. A short text about animals was input to the speech synthesizer. From the four males and the four female synthetic voices available in this system we selected the voice we considered to be the most natural sounding. This happened to be a male voice. The system was set to read at two speeds, a "normal" speech rate of 150 words per minute and a "slow" rate of 75 to 80 words per minute. Our test subjects were asked to transcribe the text they heard as faithfully as possible. The transcription was to be done in normal, orthographic English, as only two individuals were familiar with phonetic transcriptions. The text could be replayed any number of times since the subjects could not keep pace with the speed of the synthesized reading.

We were aware that knowledge of context and general understanding of the meaning of the text passage would compel a listener to "correct" his or her transcription. Thus, a word that was heard as "mat" or "pet" would be corrected to spell "bat" just because it made better sense in the given context. Insofar as substitutions

based on linguistic or extralinguistic knowledge are a feature of any listening situation, we did not feel that this impaired our evaluation.

The test population was also asked to rate the voice and reading quality of the system on an informal scale of 1 (poor) to 5 (excellent) and jot down any comment that came to mind while they participated in the evaluation.

Text. We selected a short, non-technical text about bats from National Geographic with a fairly representative sample of phoneme and phoneme combinations for the evaluation. No attempt was made to test the speech synthesizer with homographs that were not at the same time homonyms, e.g. "lead" (the metal) and "lead" (to guide). The text is given below.

> *"Millions of bats, which create the largest colonies of any mammal, have already been buried by this practice or been forced to seek shelter elsewhere. This dilemma is becoming more acute, because many states, spurred by human accidents in such mines, have stepped up the closure rate. To protect both people and bats, over the past five years more than a hundred sturdy gates have been constructed at mine entrances, allowing bats to pass through but keeping people out."*

Test population and anticipated results. Ten subjects participated in this evaluation of *DECTalk.* With the exception of one native male, they were the same individuals who participated in our *Kurzweil VOICE for Windows* evaluation. We did not expect any differences along the male/female axis but expected speakers of English as a second language—depending on their proficiency level—to have problems understanding the speech synthesizer's dialect.

Table 11
Subjects for *DECTalk* by sex and native versus nonnative command of English

	native	nonnative	total
male	2	2	4
female	3	3	6
total	5	5	10

Sample transcriptions. The transcriptions produced by two nonnative speakers are shown in Table 12. Subject r has spoken English since childhood and uses English at home and at work. Subject c learned English as an adult and uses his/her native language at home.

Table 12
***DECTalk* Transcription Samples**

Subject	Transcription
r	Millions of bats, which create the largest colonies of any mammal, have already been buried by this practice ? *in forest* ? to seek shelter elsewhere. This dilemma is becoming more acute, because many states, spurred by human accidents in such mines, have stepped up the closure rate. To protect both people and bats, over the past five years more than *130* gates have been constructed at mine entrances, allowing bats to pass through but keeping people out.
c	Millions of *mats* which create the largest colonies of any *memo* have already been *varied* by this practice elsewhere. This delema is becoming a cute, *they cause* many *mats* spurred by human *egg* in such mines have *step* up the closure rate. To protect both people and *mats* over the past five years, more *than a hundred thirty gaps* have been constructed at *mines entrance* allowing *mats* to pass through but keeping people out.

Results. We found one recurrent transcription error, namely "one hundred **thirty** gates" instead of "one hundred **sturdy** gates." This suggests a less-than-perfect differentiation between sibilants in *DECTalk.* Instead of "**stepped** up the closure rate," on three occasions the transcription was "**kept** up the closure rate".

Errors such as "pets" or "mats" or even "gaps" instead of "bats" that surfaced in the transcriptions of nonnative speakers of English could be due to poor synthesis or poor recognition on the part of the listeners. However, given that stops are defined by their surrounding sounds—all stops are articulatorily and acoustically alike for the closure (silence) and the release of the closure (plosive burst) segments—and given that the surrounding vowels were transcribed correctly, these errors reflect more on the competence of the transcriber than the precision of the system.

The voice quality and the reading quality of *DECTalk* were considered acceptable—the average rating according to our informal rating scale was 3 (good)—though by no means a match for a human voice. Intonation was found to be too flat. In slow mode there was too little separation between individual words.

Evaluation of DECTalk. The synthesized speech produced by *DECTalk* is somewhat monotonous and unnatural compared to human speech because it lacks the prosodic features that characterize human speech. This does not, however, seem to be an obstacle to understanding, judging from the transcriptions of our test subjects.

If understanding synthesized speech depends less on the quality or naturalness of the speech than language proficiency, nonnative speakers should produce errors similar to the ones described above when they are asked to transcribe a text read by a human. Due to time constraints, we did not do a comparison study that substituted human reading for the synthesized speech. A larger test population and a wider array of proficiency levels would show whether *DECTalk* could be used reliably to measure language proficiency for both native and nonnative speakers of English, in addition to the uses anticipated in section Telephony-based interactive voice response applications.

Conclusions and Recommendations

Based on the outcomes presented in this paper, it is clear that state-of-the-art commercial technology for speech recognition is not yet ready for high stakes applications like computer-based examinations. At the same time this study includes information about when such technology might be available.

Five years ago there was no PC-based technology for speech recognition or speech production applications. This was because speech technology was less developed and less computing power was available on PC-based platforms. The existence of current systems that perform at 70% to 80% accuracy (within that time span) is encouraging. If desktop computing capacity continues to evolve at the same rate, one could expect tremendous capacities for computing that would allow very large vocabularies to be stored and processed at great speed.

Since the last 20% to 30% of the problems with the speech technology software may be the most challenging, it is difficult to predict with absolute certainty when speech technology will be ready for operational use. Because updates of speech technology software are continuous, more studies using updated software could contribute to further knowledge about speech technology readiness. If the technology increases at the same rate that it has over the past five years, perhaps it can be used operationally within the next five years. Based on the results of the study, we recommend exploring the most current technology by continuing to develop prototype applications of speech-based testing applications. This prototyping should include applications that use both speech recognition and speech production technologies. In an effort to continue this study we recommend creating a series of items based on the current generation of speech recognition technologies. Because this technology functions well with limited vocabulary, these items should be designed to make use of this limitation. For instance, it was suggested by a reviewer that one could design an experimental prototype in which examinees had 10 minutes to train on a speech recognition system, using vocabulary from the test item. This type of task would seem to be a natural step in the evolution of computer-based items.

In conjunction with the development of items that required speech recognition, we also recommend a series of items using computer-based speech production as the means to deliver instructions and the stimulus for the items. We would deploy these prototypes on a larger scale in order to measure the success of using this technology.

References

AT&T to test speech recognition for networked customer telephony applications. (1995, August). *Voice News, 15 (80),* p. 1.

Faulkner Technical Reports, 1. (1994, November).

Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38(1),* pp. 35-45.

Pelton, Gordon E. (1993). *Voice processing.* New York: McGraw Hill.

Rabiner, L. A. (1995, March/April). Toward vision 2001: Voice and audio processing considerations, *AT&T Technical Journal,* pp. 4-13.

Tuttle, M. D. (1995). Saving North America's beleaguered bats. *National Geographic, 188,* pp. 36-57.

Appendix

Installation notes for Kurzweil VOICE for Windows, DragonDictate, and DECTalk

Installation notes for *DECTalk*

Installing *DECTalk* was very easy. Once the interface card was inserted, the driver software for *DECTalk* was installed and self-configured. The longest part of the installation was deciding which of the voices would be best suited for the purposes of testing the package. Total installation was about one hour, including setting up a process by which the test subjects could listen to the speech samples.

Installation notes for *Kurzweil VOICE for Windows*

Installing the MWave Communications Interface board along with the Kurzweil software was very easy, although there was a conflict between the mouse and the interface board. Since this conflict involved the modem capabilities of the interface, we chose to ignore the problem as the Kurzweil software does not require the use of the modem hardware built into the interface card. This did not have any adverse affects.

Once the Interface card had been installed into the computer, the software drivers for the card were installed, followed by the Kurzweil software itself. The total installation time (including time to isolate the aforementioned conflict) was about two hours, which included testing the software by enrolling its first user.

Installation notes for *DragonDictate*

The installation of the hardware and software for *DragonDictate* went very smoothly. Initially we installed *DragonDictate* with a Sound Blaster compatible interface called Port-Able Sound Plus. We discovered that, while Port-Able is Sound Blaster compatible, it is not *DragonDictate* compatible. After discussing this with *DragonDictate* support personnel, we de-installed the Port-Able Sound Plus interface and installed a Sound Blaster Interface card.

Following the installation of the interface card, we installed the driver software, then the *DragonDictate* software. The interface card driver software was self-configuring, which means that the installation program set up the parameters to utilize the hardware for the machine we were using in the best way. The installation of the *DragonDictate* software itself was very simple. The installer was asked whether he or she wants to preload everything from the floppies or load a minimum system and be prepared to load floppies later. The software recommends that if only one or two people are going to use the system, the minimum load choice should be selected. We recommend that no matter how many users are expected, the entire system be loaded at once to forestall the need to provide floppies at an inopportune time.

The total time to install *DragonDictate* was about 45 minutes.

Test of English as a Foreign Language
P.O. Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 609-771-7100
E-mail: toefl@ets.org
Web site: http://www.toefl.org

58713 • Y99M.5 • Printed in U.S.A.
I.N. 253714