Technology Evaluation ■

Continuous Speech Recognition for Clinicians

Atif Zafar, MD, J. Marc Overhage, MD, Clement J. McDonald, MD

Abstract The current generation of continuous speech recognition systems claims to offer high accuracy (greater than 95 percent) speech recognition at natural speech rates (150 words per minute) on low-cost (under $2000) platforms. This paper presents a state-of-the-technology summary, along with insights the authors have gained through testing one such product extensively and other products superficially.

The authors have identified a number of issues that are important in managing accuracy and usability. First, for efficient recognition users must start with a dictionary containing the phonetic spellings of all words they anticipate using. The authors dictated 50 discharge summaries using one inexpensive internal medicine dictionary ($30) and found that they needed to add an additional 400 terms to get recognition rates of 98 percent. However, if they used either of two more expensive and extensive commercial medical vocabularies ($349 and $695), they did not need to add terms to get a 98 percent recognition rate. Second, users must speak clearly and continuously, distinctly pronouncing all syllables. Users must also correct errors as they occur, because accuracy improves with error correction by at least 5 percent over two weeks. Users may find it difficult to train the system to recognize certain terms, regardless of the amount of training, and appropriate substitutions must be created. For example, the authors had to substitute ''twice a day'' for ''bid'' when using the less expensive dictionary, but not when using the other two dictionaries. From trials they conducted in settings ranging from an emergency room to hospital wards and clinicians' offices, they learned that ambient noise has minimal effect. Finally, they found that a minimal ''usable'' hardware configuration (which keeps up with dictation) comprises a 300-MHz Pentium processor with 128 MB of RAM and a ''speech quality'' sound card (e.g., SoundBlaster, $99). Anything less powerful will result in the system lagging behind the speaking rate.

The authors obtained 97 percent accuracy with just 30 minutes of training when using the latest edition of one of the speech recognition systems supplemented by a commercial medical dictionary. This technology has advanced considerably in recent years and is now a serious contender to replace some or all of the increasingly expensive alternative methods of dictation with human transcription.

■ **JAMIA.** 1999;6:195–204.

Affiliation of the authors: Indiana University, Regenstrief Institute for Health Care, Indianapolis, Indiana.

This work was supported in part by contracts N01-LM43510 and N01-6-3546 from the National Library of Medicine.

Correspondence and reprints: Atif Zafar, MD, Regenstrief Institute for Health Care, 1001 West 10th Street, RHC 5th Floor, Indianapolis, IN 46202-2859.
e-mail: ⟨zafar_a@regenstrief.iupui.edu⟩.

Received for publication: 10/12/98; accepted for publication: 1/11/99.

Medical informaticians have struggled with capturing physician-generated clinical data for a quarter century. A variety of approaches, ranging from direct typing[1–4] to menu-[5,6] and macro-based clinical note generation, have been studied and are being sold commercially (Table 1 provides a sampling of such vendors). Most physicians prefer dictation because it is simple, familiar, and fast. Furthermore, transcribed notes are suitable for entry into a computerized medical record. However, manual transcription incurs delays of hours or days and is expensive. Transcription

Table 1 ■

Vendors of Menu- and Macro-based Clinical Note-writing Systems

Product Name	Publisher	Phone Number, Fax Number, and Web Site Address
Oceania Notes WAVE EMR	Oceania, Inc. 3145 Porter Drive, Suite 103 Palo Alto, CA 94304	Tel: 888 4 OCEANIA Fax: 650 493 2202 www.oceania.com
Logician	MedicaLogic 20500 NW Evergreen Parkway Hillsboro, OR 97124	Tel: 503 531 7000 Fax: 503 531 7001 www.medicalogic.com
HealthPoint	HealthMatics 1100 Crescent Green, No. 210 Cary, NC 27511	Tel: 800 452 9653 Fax: 919 379 2200 www.healthpoint.com
ChartNote	Datamedic 95 Sawyer Road, Suite 200 Waltham, MA 02154	Tel: 781 788 4800 Fax: 781 736 0129 www.datamedic.com
Doctor's Office	PEN Knowledge, Inc. 1075 13th Street South Birmingham, AL 35205	Tel: 205 934 3718 Fax: 205 975 6493 e-mail: bcouncil@penkno.com
MediView	Physician Computer Network, Inc. 1200 The American Road Morris Plains, NJ 07950	Tel: 201 934 9333 Fax: 201 934 5538
Practice Partner	Physician Micro Systems, Inc. 2033 6th Avenue Seattle, WA 98121	Tel: 206 441 8490 Fax: 206 441 8915 www.pmsi.com

services can cost as much as 15¢ per line, or $8 per page.[7] The Indiana University Department of Medicine spends nearly $500,000 on transcription per year at one hospital. Recent advances in automatic speech recognition technology could alleviate many burdensome aspects of dictation.

Early studies of the use of voice-input devices produced disappointing results. In 1981, one researcher reported that voice entry required four times as much time as a menu selection system that had previously been used.[8] These first-generation speech recognition systems required users to pause . . . between . . . each . . . word for 200 ms and could "understand," at most, 30 to 40 words per minute, compared with natural dictation rates of 160 to 250 words per minute.[9] In addition, early systems had limited vocabularies, high error rates, and were slower than typing; consequently, they were not readily accepted. Researchers at Stanford found the early (1994) continuous speech systems to be lacking as well.[10]

Advances in speech recognition technology through 1998 produced systems that can understand continuous speech, operate in real time, run on commodity PCs, and produce more accurate results.[11] In this report, we describe this new technology and discuss its potential and limitations.

Background

Speech recognition units comprised three subsystems. A *microphone* acts as a signal transducer, converting the sound generated by the user's speech into electrical signals. A *sound card* subsequently digitizes the electrical signal. It samples the signal at various rates (typically, 6,000 to 20,000 samples/sec), creating a series of decimal numbers representing the intensity of the sound at each time point. The *speech engine* software, which behaves like a transcriptionist, then converts these data into text words.

We can express words as combinations of basic speech sounds, called phonemes. English dictionaries have symbols for each phoneme and use these symbols to describe word pronunciation. Phonemes are classified into vowels and consonants by the differences in their waveforms and vocalization techniques. We articulate vowels by arranging our vocal anatomy into relatively fixed configurations and blowing air across the vocal cords.* As the cords vibrate, a train of air impulses is injected into the vocal tract, resonating at specific frequencies, called formants.† Because of cord vibration, the waveforms of vowels show periodic behavior, i.e., they consist of repeating units of a basic waveform.‡ The rate at which these units repeat is called the pitch period.‡ We pronounce some consonants by forming constrictions in the vocal tract using the tongue and other muscles, causing momentary pauses in the speech signal, and then expelling air as if pronouncing a vowel.* The waveforms of these consonants consist, therefore, of short pauses, noted as dips in the amplitude of the speech signal. Speech that is unvoiced (like a cough or a breath) does not exhibit periodicity, which helps distinguish noise from phonemes.‡

Speech recognition is typically a two-stage process. The speech recognition system initially determines the general location of phonemes and their waveform characteristics using a process called feature extraction. It then uses pattern recognition techniques to identify the phonemes, and maps these phonemes into words.

Initially, speech recognition systems partition the continuous speech signal into equally spaced units of 10 to 20 msec, called frames.[12,13] The system then estimates speech parameters, such as the pitch period and formant frequencies for each frame. A common intermediary step for analysis of frames is to generate the power spectrum. Figure 1 depicts the raw speech waveform and power spectrum for the phrase "free speech." The dark bands on the power spectrum correspond to the formant frequencies, identifying the vowels. Notice that the vowel sound *ee* in the words *free* and *speech* has similar formant frequencies and waveform morphology in the two words. Consonants, on the other hand, are identified by the relative dips in signal amplitude. Notice that when the consonant sounds *p* and *c* are uttered, the amplitude of the waveform declines sharply. Notice, too, that the areas corresponding to the dashed arrows under the boxed word "NOISE" in the figure (unvoiced speech) show erratic waveforms, without any definite pattern.

The transformation of speech signals into a power spectrum makes it easier to identify the locations of vowels, consonants, and noise. The end result of the feature extraction is a feature vector, a set of 15 to 20 numbers that best represent a frame and are insensitive to artifacts such as noise and speaker variability.

The first step in the word recognition stage is to find the boundaries of phonemes. A rapid change in amplitude suggests such a boundary, although it is not the only criterion for boundary transitions.[14] The spectral characteristics (formant frequencies and pitch period) determined during feature extraction provide other clues to the location of boundaries. Often phonemes span multiple frames, although they may also lie entirely within one frame. So the next stage in word recognition is to combine into a segment[13] successive frames that correspond to the same phoneme. These segments, which are of differing lengths, correspond roughly to individual phonemes.

The next stage, called phoneme recognition, classifies each segment as a particular phoneme. A variety of algorithms are used to accomplish this task. One of the more common algorithms employs hidden Markov models (HMMs),[12,13] a statistical structure that encodes the probability of a sequence of events. In speech recognition, HMMs carry the probability of occurrence of each possible sequence (usually a triplet) of phonemes.[7,8] To identify the phoneme represented by a segment, we look at the waveform of the segment. Speech system designers know that the waveform of a phoneme looks different depending on the context in which it is uttered (i.e., because the waveforms of neighboring phonemes blend with it in a continuous manner).[12,13] Speech engines, therefore, inspect a segment and its two immediate neighbors as a triplet. Speech engine designers consider all possible phonemes that can lie adjacent to a given phoneme, and they store the probability of the resulting sequences (of phonetic waveforms) in HMMs. During recognition, each staggered triplet of feature vectors ((1,2,3)(2,3,4)(3,4,5) etc.) is compared with the data stored in every HMM, and a probability is obtained from each comparison. The HMM that produces the highest probability identifies the phoneme.[12,13] This process generates a temporal sequence of identified phonemes that correspond to the original speech signal.

The final stage of recognition maps the phonetic sequence into words. This stage requires a phonetic dictionary, listing the phonetic spelling of all words

*From "Articulatory Phonetics," available at http://forte.sbs.umass.edu/~berthier/ArticPhonetics.html and http://wwwdsp.rice.edu/~akira/digitalbb/formants.html.

†From "Class Notes in Articulatory Phonetics," available at http://www.umanitoba.ca/faculties/arts/linguistics/russell/138/notes.htm, and from "Speech Visualization Tutorial," available at http://lethe.leeds.ac.uk/research/cogn/speechlab/tutorial/index.html.

‡From "The Basic Properties of Speech," by Jason Woodard, available at http://www-mobile.ecs.soton.ac.uk/speech_

codecs/.

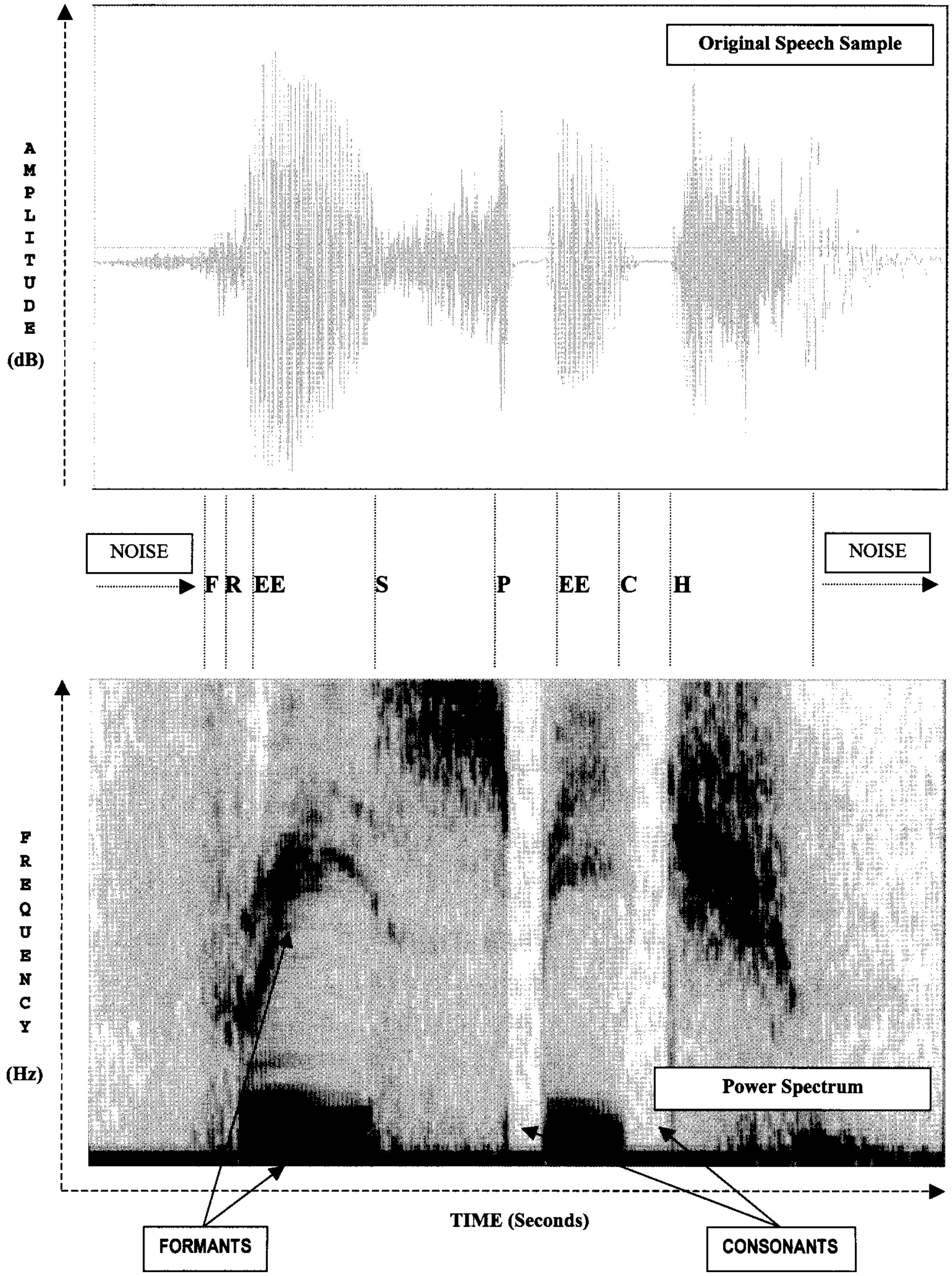

Figure 1 Spectral analysis of the words "free speech" as spoken by an author (A.Z.): *top*, the raw speech waveform; *bottom*, the power spectrum. Notice how the areas that correspond to the phoneme *ee* look similar and generate two resonance frequencies (formants) in the power spectrum. Also notice how the consonant sounds *p* and *c* produce a relative pause in the power spectrum.

that the speech engine is designed to understand, and a language model, listing the probabilities of specific sequences of words. Phonetic dictionaries are needed because English spellings alone do not define pronunciation. For example, the *f* in *freedom* represents the same phoneme as the *ph* in *phonetic* or the *gh* in *tough*. Furthermore, the *e* in *late* represents no sound at all, while the *a* in *late* represents the same phoneme as the *ei* in *weight* and appears identical in standard text to the completely different phoneme in *cat*. Language models, on the other hand, help pick out the correct words from context. For example, consider the phrase ***two** days is **too** long a wait **to** get back a lab result*. Each of words in boldface is a homonym represented by the same sequence of phonemes. However, without a language model, speech engines cannot determine which word is intended. The language model would list the sequence *two days* as being more likely than either *to days* or *too days* and accordingly choose it as the correct entry.

Table 2 ■

Comparison of the Features of the Most Popular Continuous Speech Recognition Systems

	DragonSystems' Naturally Speaking	IBM ViaVoice Gold	Phillips SpeechMagic
"Active" vocabulary	30,000–55,000 words	22,000–64,000 words	64,000 words
Speed	100–160+ wpm	125 wpm	Batch mode
Synchronous correction	Yes	Yes	Yes
Correction by voice only	Yes	No	Yes
Document navigation	Yes	No	Yes*
On-the-spot vocabulary builder	Yes	Yes	No
Training time	20–60 minutes	A few minutes	10+ minutes
Peak accuracy (per vendor)	98%	97%	100%
Speaker independent	No	Yes†	No
Speaker adaptive	Yes	Yes	Yes
Learns words from input documents	Yes	Yes	No
Facility for "macros"	Yes	Yes	No
Dictate directly into other applications	Yes	Yes	Yes‡
Software developers kit available	Yes	Yes	Yes
Hardware requirements (Windows 95)	Pentium 133, 32-MB RAM	Pentium 150 MMX, 32-MB RAM	Pentium 166, 64-MB RAM
Software price	$100–$695	$100	$6000

*The SpeechMagic system is a batch-mode-only system. Thus, documents cannot be navigated at dictation time because they have not yet been created.

†IBM supplies a "user wizard" that prompts the user to select from preset voice profiles. This is not true speaker independence, and such functionality also exists for the DragonSystems package. The user has to spend about five minutes with this wizard at the time of initial enrollment. This may have changed with ViaVoice 98.

‡Using the software toolkit the user can create hooks to send dictations anywhere. This is also true for the other two applications.

Methods

Industry reports suggest that the continuous speech products on the market have similar performances.[15] Table 2 compares the feature sets of the existing products, and Table 3 lists contact information for the vendors of these products. We report our experience with the NaturallySpeaking speech understanding software (DragonSystems, Inc.) that runs on Windows 95, 98, and NT. We performed trials of two different versions of this software. Trial 1 used DragonSystems' Personal Edition (v2.0), which comes with a built-in vocabulary of 30,000 words. We tested this system with a minimally configured computer as recommended by DragonSystems (133-MHz processor with 32 MB of RAM, a speech-quality sound card, and noise-canceling microphone) as well as a faster machine (233 MHz processor). We used a supplemental dictionary called Medi-Terms (PCP Associates), which is an 8,000-word dictionary containing internal medicine terms. This dictionary costs about $30. Additional dictionaries are available from PCP Associates for commonly used medications, surgical specialties, family practice, obstetrics and gynecology, and pediatrics, each of which is priced at $30. The Medi-Terms dictionary is also available for IBM's ViaVoice speech

Table 3 ■

Vendors of Continuous Speech Products and Dictionaries

Product Name	Publisher or Manufacturer	Phone Number, Fax Number, and Web Site
NaturallySpeaking Preferred, Professional	DragonSystems, Inc. 320 Nevada Street Newton, MA 02160 USA	Tel: 617 965 5200 Fax: 617 527 0372 www.dragonsys.com
ViaVoice	IBM 1133 Westchester Avenue White Plains, NY 10604	Tel: 800 IBM-4YOU www.software.ibm.com/is/voicetype
SpeechMagic	Philips Speech Processing 64 Perimeter Center East, 6th Floor Atlanta, GA 30346	Tel: 770 821 2400 Fax: 770 821 3687 www.speech.be.philips.com
Medi-Terms	PCP Associates 830 Potomac Circle, Suite 150 Aurora, CO 80011	Tel: 303 360 3239 Fax: 303 360 3514 homel.gte.net/kaicher/medterms.htm
Internal medicine contexts	KorTeam International, Inc. 777 Palomar Avenue Sunnyvale, CA 94086	Tel: 408 733 7888 Fax: 408 733 9888 www.korteam.com

recognition system. We did not include PCP Associates' commonly used medications dictionary in this trial.

Over three weeks, we dictated 50 discharge summaries from various internal medicine services. The average length of the dictations was 1.5 pages of single-spaced typed text, or roughly 800 words. We performed the initial training in a relatively quiet office environment, and the testing was then performed in a busy ward setting and in the physicians' workroom in the emergency room.

In Trial 2 we tested the same discharge summaries using the latest version of NaturallySpeaking software (Professional Edition, v3.0), which supports a vocabulary of 60,000 words and incorporates a new language model called BestMatch. The vendor's minimal hardware recommendations for this technology include a 200-MHz Pentium processor with 64 MB of RAM, along with a speech-quality microphone. The new system boasts improved accuracy (20 percent) when using the BestMatch technology. In this trial we used a large and more sophisticated supplemental vocabulary called ClinicallySpeaking, from KorTeam International ($695), which not only contains internal medicine terms but also a sophisticated language model. KorTeam also supplies dictionaries for other specialties, such as general surgery, cardiology, neurology, otolaryngology, and orthopedics. KorTeam's dictionaries also work with IBM's ViaVoice98 and the Phillips SpeechMagic systems. We tested the NaturallySpeaking Professional software on a 233-MHz system as well as a 300-MHz Celeron A system. These machines were equipped with 128 MB of RAM and the same sound card and microphone as in Trial 1.

Results

Trial 1

The MediTerms dictionary (8,000 words), although inexpensive, did not provide about 10 percent (400) of the terms we required to dictate 50 discharge summaries. Many of these terms were commonly used abbreviations (such as HEENT) and medication names. Since the system matches voiced words to words in its dictionary, a word absent from the dictionary is replaced by one or more similarly sounding words (phonetic analogs) that *are* present in the dictionary. For example, lack of a vocabulary for medications caused the system to transcribe "put him on heparin and nitro paste" into "put him on Hackman and mitral paste" and "Lasix" into "lay 6." Even when a term was present, its correct tense was often missing. So a phrase such as "the pain was *radiating* into her left arm" was translated by the computer into "the pain was *radiate* into her left arm." However, words can be added to the DragonSystems (as well as the

IBM ViaVoice and Phillips SpeechMagic) vocabulary, and as we added those missing terms, recognition improved. After we added the 400 new terms we were able to get accuracy as high as 98 percent. However, such accuracy required three weeks of persistent use and supplementation of the vocabulary as described. Moreover, the system was unable to understand some short phrases, such as "bid," "tid," and "qid," even after repeated training with the utterances "bee eye dee," "tee eye dee," and "queue eye dee." We had to substitute "once a day," "twice a day," and "four times a day" to get accurate recognition. In terms of the speed of the hardware, we learned that the first test system (133-MHz Pentium processor) was not fast enough. The computer would quickly fall behind in its translation after one or two sentences and get progressively further behind after that.

Trial 2

We retested some of the dictations using Professional Edition (v3.0) and the ClinicallySpeaking dictionary. We encountered none of the above problems with missing abbreviations and terms, or with recognition of such phrases as "bid" and "tid," and our accuracy was 98 percent after just 30 minutes of training. This is a major improvement over results of our first trial, which we attribute to the improved language model within the BestMatch technology and the more comprehensive ClinicallySpeaking dictionary. Thus, the adequacy of the dictionary and language model is a major determinant of recognition accuracy. In this trial with Professional Edition, a 233-MHz system as recommended by the vendor was not fast enough. We had to use a 300-MHz system to keep up with the dictation rate of about 150 words per minute.

In both trials we had to speak clearly and continuously and not pause in the middle of a sentence. The computer tends to "hear" extra words during periods of silence and either mis-recognizes the last word before the pause or adds extra words in place of the pause. Clear speech is exceptionally important when short words are spoken (such as *and*, *if*, and *the*). For example, with slight slurring of the phrase "this is a test of slurred speech," the computer heard "this isn't best of store speech." Similarly, a pause in the sentence "this is an example ⟨pause⟩ of medical dictation" was transcribed as "this is an example cough medical patient." The system is less sensitive to slurring when longer words (especially medical terms) are dictated, because they have fewer phonetic analogues. Similarly, pauses at the end of a sentence are less problematic than pauses in the middle. These problems occurred with both the Personal Edition and the latest Professional Edition (v3.0) of the NaturallySpeaking software. We also found that turning off the microphone during pauses eliminates most of the problem with pauses and improves the accuracy. So users may want to obtain a microphone with an on/off switch.

We, along with other researchers,[16] have found it best not to read the on-screen speech translation while speaking the sentence, because the system revises its interpretation as it goes (as each new word helps it better understand the previous word). Watching these changes is distracting and slows dictation. On the other hand, we (and others[15]) found it best to correct errors in the clause or sentence just completed, rather than waiting to finish the dictation, because the errors are more immediately recognizable and correcting them is easier at that point.

We also found that ambient noise (within reason) had no real effect on the accuracy. We trained the system in a relatively quiet office area and then tested it in a hospital ward, emergency room, and office. Neither the accuracy nor the speed of understanding was different in any of the three settings. So, in contrast to some discrete speech products, user's systems do not have to be retrained in the environment of intended use. In our test environments the ambient noise came from overhead paging systems, air conditioners making loud "whirring" sounds, pagers going off, residents talking in the background, the user coughing, and doors slamming shut. For example, in the presence of loud overhead announcements, the computer did add the word *to* to the phrase "we will place the patient on heparin" to yield "we will to place the patient on heparin."

Microphone placement can affect accuracy, because microphones have optimal frequency response characteristics based on their distance from the sound source.[17] Thus, the microphone should be placed at a consistent distance from use to use, ideally an inch away from the user's mouth.

The voice training file, individualized for each speaker, is large (10 MB per speaker), taking 45 seconds to load on a 13-MHz PC with a 10-msec disk access time. In a networked environment, the loading time for the speech file will be two to four times slower.

Finally, user's systems must be equipped with speech-quality sound hardware. Many laptop computers do not provide such hardware, and users need to check with their speech system vendors for acceptable alternatives.

Discussion

Voice recognition systems can provide many benefits for medical practitioners. First, they produce legible documents in electronic form that are suitable for use with a computerized medical record. Second, those documents are produced immediately, without the delay inherent with transcription services. Finally, the out-of-pocket cost per document is minuscule compared with what transcription services charge. However, speech recognition technology will consume additional user's time compared with manual transcription, especially in the early months of usage.

Before users implement such systems, they should consider a number of additional issues. First, no system is 100 percent accurate. Even after optimizing the dictionary and training the system, users will encounter errors and have to correct them. They must be willing to spend the extra time required for correcting these errors. However, for one of us with 10 years of typing experience, voice input, even after taking into account the error correction time, was significantly faster than keyboard input. With the 300-MHz machine and the professional edition of the software, we were able to dictate at 150 words per minute. Second, if multiple users share a workstation, they will face some delays, because the system must reload the speech files for each new speaker, which takes one to two minutes. Because of this, we recommend that each user have a dedicated workstation for each clinic session, so that in a busy clinic time is not wasted while speech data are loaded.

Some findings from our initial trials were unexpected. Ambient noise did not seem to have an appreciable impact on error rate, which means that users can place their speech data at multiple sites, without the user having to "retrain" in the new environments. This is useful, for example, if a clinician dictates in the clinic and then sees a patient on the wards, where another dictation is required. As long as the users employ a similar hardware configuration (e.g., users' own portable computers), they should achieve consistent accuracy while dictating in different areas.

The system we tested is highly vulnerable to changes in a user's pronunciation. Other products have similar deficiencies, and researchers at KorTeam have demonstrated that up to 8 percent of the errors are generated by inconsistencies in how users speak into a voice recognition system (R. Hendron, unpublished internal trial data, KorTeam International, Apr 1998). If, for example, a user trained the system when feeling well and then used it later with a hoarse voice, the error rate would climb. In such a case, it would be best to retrain the system for a few minutes before dictating. When the user's voice improves, the user must perform another few minutes of training or restore a previously saved training session to revert to the original voice profile.

We found that the adequacy of the dictionary is the strongest determinant of success. For high-speed voice recognition, the dictionary must contain the words that will be used and their commonly used synonyms and tenses. Casali et al.[18] demonstrated that a 25 percent reduction in the availability of the required words leads to a twofold increase in document completion time (dictation and error correction).[18]

Medical dictionaries are available for a variety of contexts, including pathology, radiology, emergency medicine, internal medicine and its subspecialties, and the surgical specialties. User's speech system vendors can provide them with a list of the available dictionaries and information on how to purchase them. These dictionaries are well worth their cost in terms of training time and accuracy of speech understanding. The NaturallySpeaking product, like several of its competitors, also allows users to load a pretranscribed text document (perhaps one of the user's own dictations) into the system, which then determines which words in the document are *absent* from its dictionary. In this way, users can quickly optimize the dictionary to suit their dictation needs.

Readers must recognize that these systems *do not* generate structured text (i.e., text organized into categories such as diagnoses, procedures, medications, and such and coded using ICD-9-CM or another coding scheme). The output is free text, like a word processor document. However, this text is stored in electronic form and can be included in a computerized record system. Some of the systems, including the Dragon Systems product, allow users to define templates (subheadings with fields that can accept input) into which they can dictate free text. This mechanism produces semistructured text that is not coded but is organized into categories. Many speech systems also allow users to define macros (paragraphs of commonly dictated text, such as the reading of a normal chest x-ray) that can be embedded, by the saying of its name, at any point in a document. This can be a great timesaver.

Most speech systems also allow voice commands. For example, if users make a mistake in the middle of a sentence, they can say "scratch that" in NaturallySpeaking (or similar phrases in competitors' products) and the system will erase the user's last utterance.

Table 4 ■

Questions To Ask a Speech System Vendor

Usability issues:

- Is there a dictionary or language model available for my dictation context?
  - Can the system learn new words from imported documents?
  - Does the dictionary have specific abbreviations that I will use?
  - Is the language model good enough to distinguish between commonly used tenses of a word?
- How easy is it to correct errors on the fly? Do I need to use the mouse or keyboard while correcting errors, or can I use voice alone?
- Can I create templates and macros to simplify complex dictation tasks?
- Can I use voice commands to navigate a document and correct errors?
- What is the baseline machine configuration I need for the system to keep up with my dictation speed—what speed processor, how much RAM, and how fast a hard disk?
- How large are the speech files, and how long will it take to load speech files when switching between users?
- How good is the dictation accuracy if my voice changes (e.g., if I have a cold)?
  - Will I need to retrain for a short time if this happens?
  - How clearly must I speak in order for the system to understand my voice (i.e., how much slurring does it tolerate)?
- How well is the system able to deal with pauses and stray sounds? (Must I speak continuously, or can I take breaks between sentences while I think?)

System issues:

- Will I need to upgrade my system to use the speech recognition product?
  - Can I use the product on my laptop system?
  - What sound hardware and microphone do I need?
- Does the system permit dictation into other applications (e-mail systems, word processors)?
- Is there a mechanism for integrating the speech software with my medical record system?
  - Can the system import data (medication lists, laboratory reports) from other applications?
  - Can I dictate directly into a note field in my medical record system?
  - Will voice commands work inside the medical record system?
- Can the system run in a network environment?
  - Can I transport my speech files from one computer to another, or will I need to retrain on the new computer?
  - How long will it take to download speech files from a central server (i.e., how large are the speech files)?

Other voice commands exist for various purposes, such as document navigation and word training. For most commercial products, training the system to recognize new words is a simple task. Users can spell the word using their voice, as the system will recognize their pronunciation of the individual letters of the alphabet. Alternatively, users can type out the word using the keyboard and then train it. Other systems have similar functionality, and many systems allow users to define their own voice commands.

Many speech recognition systems also allow dictation into any Windows application. This can be useful when a user wants to dictate e-mail or other correspondence using a word processor or computer mail system.

Some systems, including DragonSystem NaturallySpeaking v3.0, can also accept input from external devices such hand-held tape recorders and digital dictation machines. They also accept input from .wav (digitally recorded sound) files on a user's computer. This is useful if the user wants to dictate into a handheld device and later play back the recording into the system for transcription.

In summary, Table 4 lists several criteria that should be considered when evaluating and selecting a speech recognition system.

Conclusions

Voice recognition technology has the potential to overcome one of the most significant barriers to implementing a fully computerized medical record, namely, direct capture of physician notes. To realize the cost savings from the current generation of speech technology, users must select and utilize fast hardware configurations and a comprehensive dictionary that includes the words they want to transcribe. Users must be persistent in error correction over the short term. These systems have become much more accurate and usable within the last year. We expect that, over the next few years and decades, clinical vocabularies and speech recognition algorithms will further improve, speaker independence will be achieved, and natural language understanding will ultimately make structured dictation a reality.

The authors acknowledge KorTeam International for providing trial versions of their medical dictionary (ClinicallySpeaking) and for their internal trial data reports.

References ■

1. Chin HL, Krall M. Implementation of a comprehensive computer-based patient record system in Kaiser Permanente's northwest region. MD Comput. 1997;1(1):41–5.
2. Sands DZ, Rind DM, Vieira C, Safran C. Going paperless: can it be done? Proc AMIA Annu Fall Symp. 1997:887.
3. McDonald CJ, Overhage JM, Tierney WM, et al. The Regenstrief Medical Record System 1998: a system for city-wide computing. Proc AMIA Annu Fall Symp. 1998:1114.
4. Tang PC, Boggs B, Fellencer C, et al. Northwestern Memorial Hospital CPR Recognition Award of Excellence. Proc Computer-based Patient Record Institute Symp. 1998:9–53.
5. Leming BW, Simon M, Jackson JD, Horowitz GL, Bleich HL. Advances in radiologic reporting with computerized language information processing (CLIP). Radiology. 1979; 133(2):349–53.
6. Musen MA, Wieckert KE, Miller ET, Campbell KE, Fagan LM. Development of a controlled medical terminology: knowledge acquisition and knowledge representation. Methods Inf Med. 1995;34(1–2):85–95.
7. Rosenthal DI, Chew SS, Dupuy DE, et al. Computer-based speech recognition as a replacement for medical transcription. AJR Am J Roentgenol. 1998;170:23–5.
8. Leeming BW, Porter D, Jackson JD, Bleich HL, Simon M. Computerized radiologic reporting with voice data entry. Radiology. 1981;138(3):585–8.
9. Reed R. Voice recognition for the radiology market. Top Health Records Manage. 1992;12(3):58–63.
10. Detmer WM, Shiffman S, Wyatt JC, Friedman CP, Lane CD, Fagan LM. A continuous-speech interface to a decision support system, part 2: an evaluation using a Wizard-of-Oz experimental paradigm. J Am Med Inform Assoc. 1995;2:46–57.
11. Speech recognition: finding its voice. PC Mag. Oct 20, 1998.
12. Makhoul J, Schwartz R. State of the art in continuous speech recognition. Proc Nat Acad Sci U S A. 1995;92:9956–63.
13. Comerford R, Makhoul J, Schwartz R. The voice of the computer is heard in the land (and it LISTENS too!). IEEE Spectrum. Dec 1997:39–47.
14. Weibel A, Lee KF. Readings in Speech Recognition. San Francisco, Calif.: Morgan Kauffman, 1990.
15. Poor A, Brown B. Watch what you say. PC Mag. Mar 10, 1998.
16. Schurick JM, Williges BH, Maynard JF. User feedback requirements with automatic speech recognition. Ergonomics. 1985;28(11):1543–55.
17. Nakatsu R, Suzuki Y. What does voice-processing technology support today? Proc Nat Acad Sci U S A. 1995;92: 10023–30.
18. Casali SP, Williges BH, Dryden RD. Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. Hum Factors. 1990;32(2):183–96.