Continuous Speech Recognition: What You Should Know

Save to myBoK

by Holly Clark

Is speech recognition a futuristic dream, an emerging technology, or a current reality? How vital is speech input to a computerized patient record strategy? How can you influence your organization to investigate speech recognition options that best meet its needs? There are no simple answers to these questions. This article introduces the basic concepts behind speech recognition technology -- the most promising tool to computerize the transcription process and enable electronic data entry.

Speech recognition will fulfill an important role in the CPR transition. Currently, the majority of CPR systems focus on direct data input, or point-of-care entry, by the healthcare practitioner through a number of structured data elements. It is useful to think of continuous speech recognition as another option for electronic data capture. Even the most automated systems have shortcomings when capturing narrative text or unstructured, subjective findings. When direct data entry is not feasible, continuous speech recognition technology provides an efficient method for capturing clinical information. It is another tool that information professionals can employ to meet a wide variety of end user needs.

Much lively debate is taking place within healthcare information circles regarding speech recognition. According to the 1998 HIMSS Leadership Survey, 31 percent of the respondents indicated their organizations were likely to begin using voice (speech) recognition over the next 12 months. The marketplace may be ripe, but there is confusion and skepticism about what it is and how best to use it.

History

Research and development of voice recognition began more than 30 years ago. Electrical engineers, computer scientists, speech scientists, and linguists did much of the research. Considering variations in accents, clarity, speed of speech, homophones, and grammatical and spelling conventions, the development of this technology was difficult. (See "Three Categories of Voice Technology.")

Recent advances in software and increased computer processing power paved the way for today's speech recognition programs. Early programs relied on template matching technology or "discrete speech" which required the speaker to pause between each word. The latest technology is "continuous speech." Natural or conversational speech is recorded, and phoneme recognition is used to recognize streams of speech sounds. Phonemes are the smallest units of speech sounds, similar to "atoms" of speech.

Principles of the Technology

The core technology of speech recognition is similar among industry vendors. Acoustic signals (speech) are accepted as input, and sequences of words (text) are produced as output. The system records the speaker's dictation, digitizes and formats it, then compares it to a possible word and chooses the most likely word representation. The text is produced by the computer and corrections are made by an editor for "misunderstood" choices. Complex algorithms and statistical probabilities are calculated to recognize words from an identified subject matter.

In healthcare, voice recognition systems are being developed to take advantage of the unique "language" of various medical specialties. It is important to understand the rationale for this approach. Consider the fact that the English language has several hundred thousand possible words. Certain words are used much more frequently than others, especially if the communication is about a specific topic. For example, analyzing the dictated reports (input texts) of radiologists might yield a realistic vocabulary, or lexicon, of 20,000 words. Further analysis of the input text calculates the statistical probability of many sets of two or three word groups (bigrams or trigrams) occurring together. This process provides extremely valuable data for the software to monitor and is often called the statistical language model (SLM). Word combinations that occur frequently have a high probability of being selected. The SLM dramatically enhances the system's recognition performance because it knows the likely word combinations.

Product Features

Continuous speech recognition systems can be divided into two main categories: real-time and batch processing technology. This distinction refers to the work flow process when text is presented for review. Batch systems typically record the author's speech file, then convert it to text after the report is dictated. Text review and correction is performed by the author or an editor. Real-time systems display text on the author's computer screen soon after the words are spoken. Authors can view their work and correct it as they create the report.

Choosing which system will work best for your application requires a thorough understanding of your work flow and documentation needs. For example, a physician who currently dictates a large volume of reports that are transcribed by support staff may choose the batch model since he or she is accustomed to focusing only on dictation; therefore, the work flow is unchanged. However, authors who generate their own reports (handwritten or via computer) without the assistance of support staff may benefit from a real-time recognition product because they maintain control of the entire documentation process.

Most continuous speech recognition systems require fairly high-end PC hardware. However, all the major products clearly list the hardware and operating system requirements. Before purchasing, review these requirements to determine what you will need. Remember that the listed system requirements are really minimum requirements. You may need more processing power and memory to achieve optimal levels of performance.

Vendors are adapting their products and adding features to anticipate buyers needs. Command and control features, structured reporting modules, and customized templates are available with several systems. Portability, network solutions, and ease-of-use functionality are also important considerations when choosing the best system for your application.

Factors Influencing Recognition Quality

Practical discussions about the use of speech recognition technology for dictation/transcription usually start with the standard question "Does it work?" Many skeptics are correct in their belief that systems don't always deliver the expected recognition quality results. Marketing hype, overly anxious users, and a general lack of understanding about the technology all contribute to product quality questions. In addition, numerous technical issues can affect the accuracy of speech recognition systems.

Four kinds of recognition errors can occur:

- **Substitution** -- A word can be misunderstood completely and another word recogized in its place
- **Deletion** -- Failure to notice a word
- **Insertion** -- Words that were not spoken can be inserted
- **Orthographic representation** -- Incorrectly written representation of a word (example: too, 2, to, or two)

Most recognition errors are attributable to software, hardware, or user problems. Software quality depends on many factors, including the composition of the system dictionary and the integrity of the statistical language model. Adequate hardware will improve results and recommended system requirements need to be verified. However, the greatest degree of variability rests with the end users, specifically those who dictate to the systems.

To help speech recognition systems perform at their peak:

- Enunciate and monitor voice volume levels. Computers will try to recognize mumbled and slurred words, so speak clearly. Talking too loudly or too softly also will diminish recognition

- Take full advantage of the opportunity to fine-tune your speech profile during training. The more the system has a chance to identify your vocal characteristics, the better, so opt for the longest duration of training

- Consistent microphone positioning is important. The speaker/author should check the position during enrollment and at the beginning of each dictation session. If the microphone is too close to the mouth, the risk of hisses and pops increases. A good quality microphone, if positioned properly, will also help filter out background noise

- Perform adaptation. This process allows the system to "learn" how you speak and to become more accurate over time. Some high-end professional systems adapt to the speaker's voice, speech patterns, and speaking style, allowing the system to operate at peak performance for the "adapted" author

Current Reality and Potential

Continuous speech recognition provides another data input method to further a computerized patient record initiative. It now is a viable technology for dictation/transcription applications. Recent rapid advances have occurred, and the marketplace is eager to reap the benefits of this technology. While the future is difficult to predict, it is safe to say that speech recognition will significantly change the way we communicate with computers. As speech recognition research elevates to the next level, the ultimate goal is for natural, human-like speech interaction to replace the keyboard and mouse.

Three Categories of Voice Technology

1. **Navigation or Voice Command** -- User controls software operations by giving spoken information or directions. The computer responds by retrieving data or carrying out a task.

   Telephony applications are voice-enabled telephone programs that offer improvement over the standard touch-tone pad as the data input user interface

   Voice Activated or Key Word Systems -- Earliest form of speech recognition; small vocabulary of single words or short phrases to which a computer reacts

2. **Speaker Identification and Verification** -- User's voiceprint (unique vocal characteristics) acts as a biometric security measure. This ensures that people interacting with a system are who they say they are and have authorization to access the application.

3. **Speech (Voice) Recognition** -- The most sophisticated and complex software application that enables computers to "recognize" dictated speech and "translate" it into text. Another common term is automated speech recognition (ASR).

   Discreet Systems -- Early form of speech recognition that required the speaker . . .to . . . pause . . . between each word of dictation so that the software could identify the beginnings and ends of words

   Continuous Speech Recognition -- The most recent speech technology achievement allows speakers to dictate in their natural tone, rate, and rhythm. Advances in technology and computing power makes this a viable and accurate option for automation of the dictation/transcription function

---

Holly Clark is business development manger for Voice Input Technologies in Dublin, OH. From 1987-93, she was executive director of the Ohio Health Information Management Association.

Article citation:
Clark, Holly. "Continuous Speech Recognition: What You Should Know." *Journal of AHIMA* 69, no.9 (1998): 68-69.
