Speech Technology Magazine
March/April 2005

Transcription

What's it Good for, Anyway?

by Judith Markowitz

Human-computer dialogue systems have claimed center stage in the speech industry bowinghappily to the kudos of an accepting marketplace. This is the inverse of the state of affairs that existed in the mid-1990s when PC-based dictation using statistical processing was clearly the industry's star. Shrink-wrapped general dictation products were awarded extensive shelf space in computer stores and the speech industry had visions of a speech revolution built atop boxes of Dragon NaturallySpeaking® (DNS) and ViaVoice.

Those dictation products signaled the beginning of the mainstreaming of speech but their glory faded quickly - brought down by their inability to satisfy soaring expectations of human-like performance. They failed to provide true and complete speaker independence, near-perfect accuracy, sophisticated error correction, intelligent command interpretation, and other skills possessed by good secretaries. Consequently, the bubble of expectations surrounding PC dictation burst. Since then, many speech-industry professionals have dismissed transcription as consumer desktop technology for use as an assistive technology or by an occasional physician.

Despite the prevalence of this viewpoint, transcription technology did not remain frozen in the 1990s.

Dictation

Freeform Dictation

Consumers and others can still purchase shrink-wrapped products in retail stores. Freeform dictation remains part of the arsenal of tools for people with Repetitive Stress Injuries (RSI) and computer users who want to prevent RSI [1]. Mobile users can still record freeform dictations on certain PDAs and later upload those files to computer-based speech-recognition systems. In fact, support for mobile users is part of enterprise-wide solutions for law firms (e.g., Philips' client-management software package and Dictaphone's Enterprise Express legal workstation). Furthermore, the client-server approach in ScanSoft's DNS 8 offers enterprise-level support for a more general population of mobile employees by allowing them to download their speaker models to any client tied to the enterprise that has DNS software. This population includes, among others, physicians whose work takes them to many locations within a hospital.

Structured Document Generation

Structured document generation refers to the creation of frequently-produced, professional documents that have clearly-defined sections and highly-technical language. Structured document generation using speech is beginning to find a foothold in law firms[2] , government, and other organizations that produce large quantities of standard documents.

Medical report generation remains the most well-known and successful application of structured dictation. Estimates of the medical transcription market range around $10 billion for North America alone; the promise of cost reduction, faster turnaround, and increasingly accurate technology is driving the market towards automatic speech recognition (ASR).

Medical report-generation systems were also among the earliest commercial implementations of statistical language processing. The first was implemented by Kurzweil AI in the late 1980s when discrete[3] dictation technology became accurate and powerful enough to support commercial deployment.

Today, healthcare professionals generate reports using natural, free-flowing speech. They speak into telephones, mobile devices that can upload audio files to a computer, and microphones attached to desktop/laptop computers. The systems still require enrollment and adaptation, but enrollment can be performed in the background or offline[4] . Products are tailored to the individual sub-specialties, such as radiology, family practice, and emergency medicine and are available in a spectrum of languages. Some products allow physicians to edit as well as dictate, but most utilize an approach first introduced by Philips Speech Systems which separates editing from dictation and provides separate sets of specialized tools for each. Usually, dictated material is transmitted via the Internet to a transcription service that uses ASR to generate an initial transcript and a human transcriptionist to produce a finished document. Dictaphone Corporation also applies elements of natural language understanding into its document handling. Their system examines the allergies, problems, and procedures sections of all transcripts - whether they are generated by ASR or by other means. Using standard healthcare terminology, it replaces all non-standard terminology with standard terms taken from SNOMED and other databases. The goal of this process is to minimize miscommunication in all electronic medical records that are generated.

The healthcare market is the object of a great deal of interest, both from core technology providers and their partners, who include specialty integrators from the speech industry (e.g., VoiceBrook), transcription services companies (e.g., MedQuist), and healthcare- technology IT developers (e.g., IDX). These companies are incorporating speech into widely- deployed workflow systems, such as radiology picture archiving and communications systems (PACS). Making the dictation a facet of the working environment allows physicians to dictate their reports without taking their eyes or attention away from their work. IDX, for example, is developing PACS software that not only enables radiologists to generate reports but allows them to attach short (transcribed) comments to specific portions of an image.

Furthermore, the integration of speech into electronic work environments of attorneys has encouraged law firms to re-visit ASR as a viable working tool. “It is the intermarriage of systems that has added value to the firm,” says Paul Parsons, senior partner at Greenwoods, a London law firm. “I would not have the speech recognition system if it was not fully integrated into the case management system.”[5]

Human Dialogue Transcription

To speech-industry professionals, the term “dialogue” evokes images of human-computer interaction. This is not the case for court reporters, real-time stenographers, and others who create transcripts by repeating, word-forword, the conversations, interrogations, and perorations going on around them. Some are stenotype reporters who have turned to ASR after developing RSI, others are lured by the naturalness of the technology, and a growing number use ASR to generate real-time transcriptions, often speaking faster than two hundred words per minute[6] .

Most real-time transcription employs computer- aided real-time (CART) to deliver visual representations of speech to hearing-impaired individuals. ASR CART is an alternative to stenotype CART in closed captioning of television broadcasts as well as in classrooms, legislative chambers, and meeting rooms.

Professional voice reporters and CART stenographers use specialized software designed to support their work and they input speech using a steno mask.

The steno mask was originally developed to enable court reporters to create verbatim recordings unobtrusively because it prevents others from hearing the reporter's voice. When used with ASR, the stenomask also reduces noiseinduced errors by screening out background noise.

Many of voice reporters belong to Intersteno, an international association for voice stenographers[7] , but the use of ASR to create verbatim transcriptions isn't restricted to trained professionals. It's a useful tool for anyone involved in generating textual records of human-human dialogues. I use it to transcribe recordings of my interviews with speech-industry and biometrics vendors.

Background Operations

Generating a text record of speech isn't always the object of applications that utilizes statistical speech technology. Sometimes, generating a transcript is a hidden but necessary step in the completion of another goal. Generally, such applications employ a "rough transcript" based on speaker-independent recognition of free-form input. The transcript is rough because speaker-independent recognition is not sufficiently accurate to produce high-quality transcripts. Fortunately, the transcripts it can generate satisfy the needs of audio mining and call routing.

Audio mining systems locate words, phrases, and concepts in databases of audio or audio-visual files. They use transcripts[8] to create an index of the spoken contents of those files so that the system can locate all instances of specific words or phrases in the files. When audio mining is incorporated into call center informatics it enables management to delve deeply into large numbers of calls looking for specific items or trends. The additional information can be applied to quality control, marketing, and other customer-related functions. Intelligence agencies perform similar analyses on a variety of audio sources, sometimes performing the textual conversion and searches as they receive the audio input. Audio mining is also used in legal research and litigation to examine the contents of depositions, testimony, and other audio/audio-visual sources. The use of ASR for information retrieval is familiar to attorneys. In the mid- 1990s, legal research and information suppliers, such as West Corporation and Lexis created ASR interfaces for their legal

document search and retrieval services. Those interfaces transcribed spoken input into the SQL-query format required by the search engines.

The primary object of call routing is to transfer a caller to the appropriate department or extension. When the task involves processing a query, call-routing systems often generate a rough transcript as the first step in handling the call.

The Future

This demonstrates the extent to which statistical speech processing technology is being used for dictation, human dialogue transcription, and background operations. There is also a tremendous amount of research directed at enabling the technology to achieve its full potential. I've discussed portions of that research in my recent column on machine translation[9] and other Voice Ideas columns. I plan to examine more of that research in future articles and columns as well.

---

1 Portions of this article were generated by speech recognition.

2 Law firms have used structured document-generation tools for over 10 years. Traditionally, those tools use the mouse for selecting from among options (e.g., the kind of contract to generate) and the keyboard for entry of non-standard and unique text (e.g., beneficiary names).

3 Discrete dictation technology requires speakers to pause between words/phrases. Commercial "continuous-speech" systems did not appear until the mid 1990s.

4 Enrollment using offline batch processing uses prior dictations for that speaker paired with transcripts for those dictation files.

5 From "Voice Recognition: The Power of Speech." LEGAL IT March 29, 2001 www.legalit.net .

6 Core-technology developers design and test their systems to handle speaking rates of up to 160 words per minute, which covers the range of normal speaking rates. Voice reporters generally create special "fast speech" models to support speaking rates greater than 200 words per minute. The National Verbatim Reporters Association (NVRA, the US division of Intersteno, see note #7) runs an annual competition that requires contestants to do verbatim transcripts of speech as fast as 350 words per minute.

7 Intersteno is a 50-year-old European organization that recently merged with the NVRA.

8 Nexidia's technology generates phonetic sequences rather than word-based transcripts.

9 Markowitz, Judith. (2004,November/ December) "Automating the Tower of Babel." Speech Technology Magazine 9 (6), 44.

Dr. Judith Markowitz is the technology editor of *Speech Technology Magazine* and is a leading independent analyst in the speech technology and voice biometric fields. She can be reached at (773) 769-9243 or jmarkowitz@pobox.com.