What is Automatic Speech Recognition?

Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time (Stuckless, 1994). In a nutshell, ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text.

Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although ASR technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services.

The ultimate goal of ASR research is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. Today, if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%.

Commercially available ASR systems usually require only a short period of speaker training and may successfully capture continuous speech with a large vocabulary at normal pace with a very high accuracy. Most commercial companies claim that recognition software can achieve between 98% to 99% accuracy if operated under optimal conditions. `Optimal conditions' usually assume that users: have speech characteristics which match the training data, can achieve proper speaker adaptation, and work in a clean noise environment (e.g. quiet space).

This explains why some users, especially those whose speech is heavily accented, might achieve recognition rates much lower than expected.

History of ASR Technology

The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s. Much of the early research leading to the development of speech activation and recognition technology was funded by the National Science Foundation (NSF) and the Defense Department's Defense Advanced Research Projects Agency (DARPA). Much of the initial research, performed with NSA and NSF funding, was conducted in the 1980s. (Source: Global Security.Org)

Speech recognition technology was designed initially for individuals in the disability community. For example, voice recognition can help people with musculoskeletal disabilities caused by multiple sclerosis, cerebral palsy, or arthritis achieve maximum productivity on computers.

During the early 1990s, tremendous market opportunities emerged for speech recognition computer technology. The early versions of these products were clunky and hard to use. The early language recognition systems had to make compromises: they were "tuned" to be dependent on a particular speaker, or had small vocabulary, or used a very stylized and rigid syntax. However, in the computer industry, nothing stays the same for very long and by the end of the 1990s there was a whole new crop of commercial speech recognition software packages that were easier to use and more effective than their predecessors.

In recent years, speech recognition technology has advanced to the point where it is used by millions of individuals to automatically create documents from dictation. Medical transcriptionists listen to dictated recordings made by physicians and other health care professionals and transcribe them into medical reports, correspondence, and other administrative material. An increasingly popular method utilizes speech recognition technology, which electronically translates sound into text and creates transcripts and drafts of reports. Transcripts and reports are then formatted; edited for mistakes in translation, punctuation, or grammar; and checked for consistency and any possible errors. Transcriptionists working in areas with standardized terminology, such as radiology or pathology, are more likely to encounter speech recognition technology. Use of speech recognition technology will become more widespread as the technology becomes more sophisticated.

Some voice writers produce a transcript in real time, using computer speech recognition technology. Speech recognition-enabled voice writers pursue not only court reporting careers, but also careers as closed captioners and Internet streaming text providers or caption providers.

How Does ASR Work?

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech (i.e. the microphone).

This process begins when a speaker decides what to say and actually speaks a sentence. (This is a sequence of words possibly with pauses, uh's, and um's.) The software then produces a speech wave form, which embodies the words of the sentence as well as the extraneous sounds and pauses in the spoken input. Next, the software attempts to decode the speech into the best estimate of the sentence. First it converts the speech signal into a sequence of vectors which are measured throughout the duration of the speech signal. Then, using a syntactic decoder it generates a valid sequence of representations. (Rabiner & Juang, 2004)

What is the Benefit of ASR?

There are fundamentally three major reasons why so much research and effort has gone into the problem of trying to teach machines to recognize and understand speech:

- Accessibility for the deaf and hard of hearing
- Cost reduction through automation
- Searchable text capability

What's been happening in ASR?

Aside from the scientists and technicians who are engaged in ASR research and development, most people who think about ASR underestimate its complexity. It is more than automatic text-to-speech, ASR requires fast

computers with lots of data capacity and memory--a necessary condition for complex recognition tasks, and the involvement of speech scientists, linguists, computer scientists, mathematicians, and engineers.

The search is on for ASR systems that incorporate three features: large vocabularies, continuous speech capabilities, and speaker independence. Today, there are numerous systems which incorporate these combinations.

What's ahead?

Encouraged by some innovative models, developments in ASR appear to be accelerating. The outlook is optimistic that future applications of automatic speech recognition will contribute substantially to the quality of life among deaf children and adults, and others who share their lives, as well as public and private sectors of the business community who will benefit from this technology.

References

Stuckless, R. (1983). Real-time transliteration of speech into print for hearing impaired students in regular classes. American Annals of the Deaf, 128, 619-624.

Stuckless, R. (1994). Developments in real-time speech-to-text communication for people with impaired hearing. In M. Ross(Ed.), Communication access for people with hearing loss (pp.197-226). Baltimore, MD: York Press.

Rabiner, Lawrence R. and Juang, B.H. (2004). Statistical Methods for the Recognition and Understanding of Speech. Rutgers University and the University of California, Santa Barbara; Georgia Institute of Technology, Atlanta

Docsoft Inc. has been promoting the use of Automatic Speech Recognition Technology through the marketing of our products since 2002. Visit us online at www.docsoft.com. If you are looking for more information about the usability of Docsoft:AV and Docsoft:AV Services, please contact us. We can be reached toll free at 1-877-430-3502 or at info@docsoft.com.