Automatic Speech Recognition and Access

Posted here with permission from SHHH:

Davis, C. (2001). Automatic Speech Recogntion and Access: 20 years, 20 months, or tomorrow? *Hearing Loss,* 22*(4),* p. 11-14.

Cheryl D. Davis, Ph.D.

By now you have probably heard news reports on automatic speech recognition (ASR) technology for use with computers and the claims of how it will change the world. One can write a paper, compose an e-mail message, and open programs without ever touching the keyboard. The new claims come about with the advent of continuous speech rather than discrete speech or stop speech, as it is sometimes called. In the past, one had to speak one word at a time, pausing distinctly between words. Although this was fairly easy to achieve with practice, most still found it cumbersome and undesirable.

Recently, though, several breakthroughs in the technology, such as greater memory allowing for larger dictionaries, and faster processors, have made continuous speech a reality. This means that one can speak at rates of up to 160 words per minute and have the computer automatically covert this speech to text, with error rates of 3 or 4 percent. The more you stop and correct your errors, the better the program becomes at recognizing your speech patterns. In fact, some programs advertise that with only 30 minutes of training the program to your voice, you can be up and running with ASR (in the past, 1-2 hours of initial training, that is the reading into a microphone of scripted text, was required). Add to this that the programs are coming down in price, and this is exciting news, indeed.

If the reader's reaction is anything like mine, you are half-way to the store already. You might be dreaming of having your Western Civilization professor, spouse, or boss speak into a microphone, and voila! The words appear on a computer screen as text. Unfortunately, it is not quite that straightforward. What I would like to cover here are the limitations of current speech recognition technology, and explain what might be required for it to have usefulness to individuals who are interested in using it for access to the spoken word, and not as the hands-free dictation system for which it is being developed.

ASR and Hands-free Dictation

First, it is very exciting that ASR has moved beyond discrete speech and into continuous speech. There are a variety of models and techniques that must be incorporated to ensure accurate results. (See Stuckless, 1997 or http://www.isc.rit.edu/~ewcncp/Lovejoy.html for a more in-depth treatment on modeling techniques).

To help clarify what an accomplishment this is, let's examine our speech. Spoken words are like snowflakes. No one person ever says a given word exactly the same way twice. This is not solely the result of inflection, emphasis, emotion, or being congested. When words come at the beginning of a sentence or at the end of a

sentence, they are pronounced differently. Looking at spectrograms of words, the length of time spent on a vowel is different when we are emphasizing that word than when we are not. In fact, consonants are pronounced differently depending upon the vowel that follows it. These differences appear in length of the pronunciation of the consonant, pitch, and emphasis (this is referred to as co-articulation, or the way a particular sound is influenced by other sounds near it, e.g., see, say, sow, soy). In addition, the duration of the vowel before a voiceless consonant is shorter than before a voiced consonant (e.g., beat and bead) (Mandel, 1997).

There are also issues around detecting word boundaries. We do not pronounce words as distinct entities when we are speaking. For example, the phrase ‘Coke is it’ is pronounced in the consonant-vowel combinations ‘co ki sit.’ The acoustic cues for word boundaries are very subtle, and we often rely on context to determine them. Allen (1997) gives the examples “Pulitzer Prize” and “The stuff he knows can lead to problems.” How do we know what was said was not “pullet surprise” or “The stuffy nose can lead to problems?” These are acoustically identical. There are a vast number of homophones (words that sound the same but that may be spelled differently, such as pair, pare, pear) in the English language. We are asking a computer program to not only take in the acoustical information, but to evaluate the context to determine meaning.
Programs can use a variety models to evaluate which word might logically follow another, often incorporating syntax or linguistic cues. Words are categorized as nouns, verbs, articles and so on and the rules of the English language are applied. This resolves a number of problems in word identification, especially when one is carefully composing a message or paper.

ASR and Natural Settings

When you think about the complexity of the English language, and add to that the complexities of the acoustics of spoken language, it is indeed wondrous that programs have been able to make sense of as much of our verbal communication and achieve the word accuracy that they have. However, several problems remain. Let’s shift our focus now from the controlled albeit complex examples described above to the language and dialogue that programs would be required to handle in natural settings.

Natural speech, or speech used in natural settings, is very different from the scripted written word or carefully composed dictation one might use at a computer work station. We often do not use syntactically correct sentences when speaking. Our natural speech includes many false starts, stammers, repetitions, and incomplete sentences. If you need proof of this, look at a realtime transcript of any presentation, courtroom deposition, or SHHH meeting, and you’ll realize just how human we all are when it comes to syntax. Programs are becoming better at handling these dysfluencies, but the greater the number of these dysfluencies, the greater the error rate of the transcription.

All of the examples above referenced a single speaker only. In a natural setting, such as an office or classroom, there will be multiple speakers. Many programs can be trained for several speakers, and even different languages, but the information is stored in separate dictionaries. You must let the computer know which speaker’s dictionary to use before you begin, and during that session you will be using only that dictionary. Currently, no program will recognize and accurately transcribe random speakers. Thus, you could not set up a computer with ASR next to a TV and expect to see the evening news transcribed. While many of the problems with speech variability have been overcome on a single speaker basis, the differences that multiple speakers introduce, such as variations in pronunciations, accents, and rate of speech have not yet been conquered.

This introduces the issue of other cues we pick up acoustically. These are critical to the use of ASR in natural settings. First, we understand that a dialogue is happening because we hear different voices and see different people speaking. With the exception of conference telephone calls, people rarely identify

themselves before they begin speaking. In order for ASR to be used in a natural setting with multiple speakers, some type of indication of a change in speakers would be necessary or the message would quickly become unmanageable (Stuckless, 1999).

Even when there is only one speaker, a critical issue remains. Sentence markers, such as commas, periods, question marks, and change of topic, in spoke language are communicated through pauses, inflection, and other acoustic cues. This is rarely discussed with ASR because ASR is being developed as a dictation technology, not an access technology. In dictation, one speaks punctuation. If you were to dictate the first sentence of this paragraph, you would say “new paragraph even when there is only one speaker comma a critical issue remains period”. Imagine this article without punctuation, without capitalization, without paragraph breaks. It would be extremely difficult to read and understand, but you could probably figure out a great deal of it with some study. Now imagine that instead of reading a well-scripted paper without punctuation, you are reading a transcript of someone presenting this information spontaneously, again without punctuation. Now suppose that instead of the presentation being a monologue, there are questions interspersed by two or more speakers. Finally, imagine the text appearing on a screen at 160 words per minute. In a natural setting ASR quickly breaks down. The transcript would be next to impossible to decipher even without transcription errors.

Access Uses of ASR

Now that I have described the limits of ASR, are there any situations where ASR might be used for access? Consider that notetaking or handwriting is 20 or 30 words per minute, typing is typically 40 – 60 wpm, computerized notetaking and summary systems can improve that speed to 100 – 120 wpm, and that ASR can handle speech at approximately 160 wpm. Only realtime transcription is faster, at about 300 wpm. Let’s look at what ASR can do, and be creative about fitting it to some of our needs.

In order for ASR to work, several criteria must be met. A speaker must develop a dictionary within that program. That dictionary cannot be used for other speakers because it will become contaminated. Similarly, a good microphone and quiet environment are required. A quiet environment ensures that background noise will not contaminate the dictionary. A carefully placed directional, noise-reduction microphone will also improve reception of the target speech and reduce unwanted background noise. If there will be technical language, acronyms, or jargon used during the session, that information should be read into the dictionary ahead of time. Finally, the person reading the ASR transcript will need to be skilled in English in order to overcome some of the errors that will occur in the transcript. The reader must be able to gather from context what the mistranslated word or sentence should have been, or to insert the correct homonym.

Some possible settings that immediately come to mind are situations where there is one-on-one interaction between a hearing person and a hard of hearing or deaf person. This might be a tutoring, counseling session, a meeting between employer and employee, or any other regular meeting that takes place between two people. When it is an interaction rather than a presentation, there is the opportunity for the individual to provide a few sentence markers, such as “new paragraph” or “period” (this inserts a period, a space, and capitalizes the first letter of the next word) to make the text more readable. If both individuals can see the computer screen, the speaker will have the opportunity to clarify or correct what was said, and monitor how full of text the screen is becoming. Interestingly, because these programs can accommodate a variety of languages, you might use this as a tool in working with a foreign language tutor.

What about lecture situations? One advantage of ASR is that there is no keyboard or typing noise, a complaint that is sometimes made about computerized notetaking. Because the programs can handle multiple speaker dictionaries, each teacher could train the program to his or her voice. Questions from the class could be handled in the same way that questions are handled with assistive listening devices (such as personal FM systems) are used; that is, the instructor would repeat the question into the microphone, and

then answer it. But there still would be the difficulty of the instructor not being able to correct errors during a lecture, the difficulty of reading a lecture without punctuation, and the logistics of having each instructor update the program's dictionary for various lectures.

Obviously, there are a number of logistical problems that make the use of ASR with multiple lecturers difficult or impractical. There is, however, research being conducted to determine its usefulness as a tool of access. One of these is evaluating the use of shadowing as a way around the multiple speaker problem (Stuckless, in progress; Stinson, M., Eisenberg, S., Horn, C., Larson, J., Levitt, H., Stuckless, R., 1999). With shadowing, an individual develops her own dictionary, wears a special mask with a built-in microphone (a stenographer's mask) that is connected to the computer with her speech files. She repeats what the lecturer says as fully as possible, adding sentence ending punctuation and identifying changes in speakers. Now you no longer need each instructor to spend time developing the dictionary, you only need the instructor to provide this information to the shadower ahead of time so that the vocabulary can be built into the speech dictionary. Like a notetaker or interpreter, the shadower moves from class to class with her equipment to provide access. In some ways, this type of transcript may be more readable than natural speech, as the shadower will listen to what is said, and summarize or rephrase if necessary (as in the case of false starts or incomplete sentences). Also just as with interpreters or other transcriptionists, the instructor will need to ensure that only one person speaks at a time, and follow other rules of communication that are commonly employed when communication accommodations are being made.

Another study, the Liberated Learning Project (http://www.liberatedlearning.com) is currently developing a software program that can be used with IBM Via Voice that will address some of the difficulties described above. For example, a pause in speaking will insert a hard return, making the text more readable, thus reducing the need to speak punctuation. This international consortium of university and industry partners is examining the potential of ASR in the classroom.

Finally, Sprint and Ultratec have been experimenting with the verbatim shadowing technique described above to determine the usefulness of ASR for relay operators (Coco, 2000). The communication assistant shadows (as described above) what the hearing caller is saying, and has the opportunity to correct what appears on-screen before it is sent to the deaf or hard of hearing caller. As with other applications of ASR, as long as the vocabulary is in the communication assistant's dictionary, the accuracy is quite high. However, when technical jargon is used, or when the speaker has a high rate of dysfluencies (e.g., uh, um), the accuracy is greatly reduced and may require more work on the communication assistant's part to correct it.

Summary

As can be seen from the above discussion, more and more experimentation is being done with ASR as a tool for access, creating exciting applications for hard of hearing and deaf consumers. Although ASR is promising technology, it may be many years before it will be useful in natural settings. Nonetheless, there are situations, given the right conditions, where it could be very useful in providing communication access.

Resources:

Allen, J. (1997). Applications of automatic speech recognition to natural language and conversational speech. In R. Stuckless (Ed). Frank W. Lovejoy Symposium on Applications of Automatic Speech Recognition with Deaf and Hard of Hearing People. Rochester, NY: Rochester Institute of Technology.

Coco, D. (2000). Speeding up relay services via ASR (that's automated speech recognition). Hearing Health, 16 (1), pp. 82-84.

Mandel, M. (1997). Discrete word recognition systems and beyond: Today and five-year project. In R.

Stuckless (Ed.). Frank W. Lovejoy Symposium on Applications of Automatic Speech Recognition with Deaf and Hard of Hearing People. Rochester, NY: Rochester Institute of Technology.

Stinson, M., Eisenberg, S., Horn, C., Larson, J., Levitt, H., Stuckless, R. (1999). Real-time Speech-to-Text Services: A report of the National Task Force on the Quality of Services in the Postsecondary Education of Deaf and Hard of Hearing Students. Rochester, NY: Northeast Technical Assistance Center, Rochester Institute of Technology. Or http://www.rit.edu/~netac/publication/taskforce

Stuckless, R. (1997). Frank W. Lovejoy Symposium on Applications of Automatic Speech Recognition with Deaf and Hard of Hearing People. Rochester, NY: Rochester Institute of Technology. Or http://www.isc.rit.edu/~ewcncp/Lovejoy.html
Stuckless, R. (1999). Recognition means more than just getting the words right: Beyond accuracy to readability. Speech Technology (Oct/Nov '99), pp. 30-35.

Other Related Websites

http://tap.Gallaudet.edu/speech.htm
http://www.usc.edu/ext-relations/news_service/real/real_video.html Demonstration of the Berger-Liaw Neural Network Speaker Independent Speech Recognition System
http://www.odc.state.or.us/tadoc/techcp26.htm Speech Recognition Technology by Gary Robson, Cheeta Systems

---

Direct suggestions, comments, and questions about this page to:

Cheryl D. Davis, Ph.D., Coordinator
WROCC Outreach Site at
Western Oregon University
Monmouth OR 97361
503-838-8642 (v/tty)
503-838-8228 (fax)
http://www.wou.edu/wrocc
wrocc@wou.edu

Last updated on 03FEB03.