The Growing Impact of Speech Technology on Society

Patti Price, PPRICE Speech and Language Technology

Because spoken language is pervasive, speech technology has the potential to make information more accessible to more people, in more places, more often. But our vast experience with speech also leads to frustration in interactions with automated spoken language systems when normal conversational expectations are not met.

In the sense that 'technology' is anything that did not exist before you were born, speech was perhaps the first big information technology, at least for our early ancestors --- although it's hard to imagine early parents trying to get the hang of speaking as their teenage children used speech constantly. Speech evolved with us over the past centuries as a social mechanism for exchanging information. It's a part of us in a way that writing, computers and other things we normally call technology are not. Although significant progress has been made in automating speech recognition over the past 50 years or so, the aspects of speech that make it so hard to automate arise chiefly from its social use. The frustration we feel in interacting with automated systems arises chiefly from violations of our social expectations.

What is Automatic Speech Recognition?

The goal of automatic speech recognition is to extract a string of words from the speech signal. The output does not include who is speaking (that's speaker recognition or speaker verification). The output does not include what the words mean (that's natural language understanding or related technologies). Speech recognition takes as input an audio signal and produces as output a string of words in text form representing the transcription of the speech input. In speech perception by humans, visual information is also normally used, and researchers in automated speech recognition are also exploring the use of this source of information.

Speech recognition systems in commercial use today are based on probabilistic modeling. These systems require a large set of training data. The data includes examples of transcribed text from which the models are trained. Systems typically do not perform well when the speech to be transcribed differs significantly from the training data.

State of the Art in Automatic Speech Recognition

A common question is "how good is the best speech recognition today?" Because the approach is probabilistic, this is a difficult question to answer in a general way. On the one hand, every system can have near perfect performance, depending on how the test is set up, and, depending on these conditions, one system or another may be favored. The only way to understand performance numbers is to know exactly the conditions of training and testing. And the only way to compare performance of two systems is by training and testing them in the same way. This has been the purpose over the past 20 years or so of common benchmarks used in the research world, and they have been tremendously beneficial to progress in the field. In the commercial world, because so many factors in training data and user interface may

vary, it is difficult to compare systems – perhaps the only answer there is: Is it good enough for the proposed purpose? For example, applications like Google search have taught us that systems can add great value even if we cannot measure accuracy very well. They just have to be good enough to provide added value in a given context.

Technology developers, however, need to measure improvements and typically compare performance on the same (or very similar but new) test set across many different conditions of training parameters. If the identical test set is used, there is a danger that the decisions made are too closely tied to the training and testing data and will not generalize. If the test sets are too different from each other we don't know if an increase in performance arises from changes we made in the experiment or because the new test set is more challenging or less challenging than the previous set.

Typically, performance is measured by aligning a human transcription of the speech with the computer generated transcription and counting the percentage of insertions, deletions, and substitutions. Because insertions count as errors, accuracy can be a negative number. For example, if the phrase "recognize speech" is transcribed as "wreck a nice beach", an alignment program might align 'recognize' with 'wreck' and align 'speech' with 'beach' resulting in 2 substitution errors and 2 insertion errors ('a' and 'nice') and no deletion errors, for a word error rate of 4 errors for the 2 words, or accuracy (one minus error rate) of negative 100% in this example.

State of the art systems typically do far better than that on average, but are still often far from human performance in many ways. Several factors currently constrain performance and very high performance can be achieved if one or more of these factors is controlled. For example,

- **Noise Environment.** A very good microphone, very close to the speaker's mouth in a very quiet room minimizes the noise and improves performance. Performance will decrease with lower quality audio signals and noise of various types, particularly with competing speech.
- **Speech Style.** A very careful speaker fully articulating all utterances and never hesitating or making speech errors or getting interrupted is much easier to transcribe accurately than is someone with a mouthful of mashed potatoes speaking excitedly with a friend.
- **Dialect.** Speakers who match the training data better will have higher accuracy rates than those whose speech matches less well. For example, if only one dialect region was covered in the training data, performance can be expected to be poor on those accents representing other areas. Even more difficult is the speech of nonnative speakers, since their pronunciations may differ significantly from all or most of the training data, and they may be more variable in pronunciation and less fluent in speaking.
- **Complexity.** This is really a catch all for everything else that affects performance, such as the size of the vocabulary, the typical out of vocabulary rate, how fast the recognition must be performed, what computing resources are available, and the complexity of the language modeled. For example, a large set of words with a strong bias towards 'yes' or 'no' would be less complex than the same vocabulary with each word equally likely.

Social impact (Effect of Society on Speech Technology)

The major challenges of speech recognition outlined above largely arise from the social nature of speech:

- Noise. People have evolved to share information through speech, no matter where they are, even in quite noisy conditions. Whether at a Cro-Magnon convention or at a AAAS meeting with many people talking at the same time, near the roaring surf or the roaring freeway --- people still use speech to communicate. Noise both degrades the signal and changes the way people speak. The changes people make when speaking in noisy conditions seem to help human communication, but for our current probabilistic algorithms, it is just one more source of variability that increases the difficulty of the problem.
- Speech Style. Just as we have long had social distinctions in the way we dress, we have also used speech to mark social distinctions. The acoustics of speech differ between men, women, and children -- their vocal tracts sizes differ and are shaped differently. Social class is also marked in our speech. We have formal attire and formal speeches just as we have more casual attire and casual speech. People take many short cuts in casual speech. "Did you eat yet" might sound like "Jee chet". Recovering the intent from what is left in the signal can be difficult for people but disastrous for our systems. Both the variability and the loss of information is a challenge for speech recognition. Casual speech is especially challenging because of various spontaneous speech effects that need to be detected and removed (for example, "um", "uh", repeated words and parts of words, self corrections).
- Dialect. People who speak more with each other come to speak more alike. It is one of the markers of belonging to the group, along with dress and behavior. It is said that a language is just a dialect with its own army and navy. That is, the boundaries between dialects and languages are not as firm as they seem at first glance. The variability arising from dialect differences and nonnative speakers is a challenge for our speech recognition systems.
- Complexity. People use speech to communicate in social contexts that vary in complexity. A very beginning language learner, for example, may be competent to determine which of two movie names a movie-goer might say in requesting a ticket and not be competent to negotiate a peace treaty. Our systems are closer to the first task than to the second and only perform well in very constrained situations.

Social impact (People vs. Technology)

If, as argued above, speech and language were the first information technologies, then of course speech has had an ENORMOUS social impact on people. If we restrict the question to speech technology in our lifetimes, then we come up with something less than enormous. What we'd like from speech technology is something like the 1987 visionary film made by Apple called the Knowledge Navigator:

- Human-like speech synthesis
- Easy to use interface
- Intelligent proactive assistance in finding information (without being annoying)
- Assistance in collaboration with others

Instead, more than twenty years later, we have something more like the spoof from Saturday Night Live in 2006, embodied by "Julie the operator lady", who

- Sounds a bit robotic "I'm sorry. I didn't get that" she says when someone speaks in a foreign accent,

- Annoyingly frequently asks for confirmation: "I think you said… 'dizzy dizzy dizzy'" (when it was "busy, busy, busy") and,
- Even more annoyingly, is frequently wrong).

In short, Julie is a bit deaf, not very bright, fairly autistic, yet insecure and always seeking confirmation. In her defense, AMTRAK's use of Julie (the application and voice the spoof is based on) was a big advance. According to a 2004 New York Times article (Urbina 2004), since her first appearance in April 2001, the automation 'Julie' embodies has:

- Earned an approval rating of more than 90 percent,
- Saved more than $13 million that it would have cost for humans to handle calls,
- Saved people from a possibly boring job filled with calls from frustrated people,
- Answered the calls tirelessly, patiently, perkily, consistently and to the best of her abilities.

As for the social well-being of speech technologists, however, let's just say that their jobs are secure for now since there is plenty of room for improvement! In fact, people far outperform automated speech recognition (ASR) on many (but not all) measures.

In most noise situations, humans are better than ASR. Can we attribute this superiority more to our ability to hear the sounds or to our ability to take advantage of other knowledge sources such as word frequencies or context expectations? Sroka and Braida (2005) looked at consonant-vowel-consonant syllables in order to remove from the test the higher level language processing we know humans are much better at than ASR and focus on the acoustics of speech. In these experiments, humans outperformed the ASR systems assessed in conditions of additive noise (filtered to be shaped like speech, as opposed to white noise). However, ASR and humans had very similar performance in the case of removing some low frequency portions of the signal (high-pass filtering – a change that happens typically to us as we age and gradually lose our ability to hear higher frequencies). In the case of removing some high frequency portions of the signal (low-pass filtering), the ASR systems assessed outperformed the humans. These results are not surprising since it is only relatively recently that people have had much experience with bandpass filtered speech such as telephone speech (which removes frequencies above about 4000 Hz).

In a recent 'Recognition Challenge' Cooke et al. 2010 also compared human labeling performance to six competing ASR systems and a simple baseline ASR system under various conditions of additive noise. This task was quite challenging: find key words spoken by one of two speakers saying similar sentences at the same time. On average, ASR was much worse than human performance, particularly in the noisier conditions. Although one system was rather similar to human performance in this task, it possibly was an artifact of the creation of the stimuli -- the best system (and possibly others) took advantage of the fact that the absolute gain of the target speaker was constant.

Shen et al. 2008 tried yet another approach to separate higher level language processing from the acoustic processing, one that enabled using actual spoken utterances and not just isolated syllables. They chose languages that had similar sound systems and asked native speakers of one of them (Italian) to transcribe utterances in the other two (Spanish and Japanese). In this case, a simple speech recognizer was about as accurate as the worst of the human transcribers.

Overall, humans and ASR performance are similar in that they both:

- Degrade with noise
- Degrade when faced with an unfamiliar style, dialect or nonnative speech
- Degrade with increased task perplexity (a measure of how many words compete at any time given what has preceded)

Taken together, the various comparisons made between ASR and human performance indicate that, while there is significant variability across humans and across ASR systems, humans still tend to be more adaptable in the face of various challenges such as noise, dialect variation, etc. However, it appears that the margin of difference is narrowing and, further, that there are some tasks at which ASR outperforms humans.

A jet plane can fly faster and farther than a bird, while taking advantage of some of the same aerodynamic principles. Yet a plane is much less adept than a bird at some tasks, such as landing on a telephone wire. Similarly, ASR is far superior to people in some respects (patience, consistency, keeping track of long lists and large data structures) and far inferior in other respects (flexibility, adaptability, making inferences, etc.). Our job as speech technologists is to improve speech capabilities. Our job as applications developers is to understand what the technology can and cannot do, and to find interesting and useful things that can be managed in the current state – be it fully automated or involving collaboration between humans and computers.

Progress, Challenges, and Predictions

Progress. It is sometimes difficult to gauge progress from up close since changes may be very gradual. Based on an informal survey of speech researchers and looking back over the past decade, it appears that the major progress has involved taking advantage of faster, cheaper, larger memory and in finding useful applications. In particular, progress during the last decade includes:

- Further elaboration of statistical pattern matching schemes, and the creation of larger and more detailed modeling,
- The elaboration of standards in speech technology making application development easier,
- Applications that move beyond dictation and the automated call center, for example: transcription of video and voice messages, and searching audio streams for key words or phrases,
- The coverage of many more languages and speech companies located in many more countries, and
- Unsupervised training. Unsupervised training has been explored in recent years in research laboratories, but a striking deployment has been Google's GOOG-411 application. This system has no fallback operator in the loop and is self-adapting --- constantly using recognized utterances and its own estimate of confidence in those transcriptions to update its own training. Although supervised training may in general be better than unsupervised training, when very large amounts of data are available without the cost of supervision, it is obviously worth exploring unsupervised training.

Challenges. Given the list above, one might reasonably ask: where is the science? The developments made have largely involved taking advantage of existing technology for new purposes, using more and more data with more detailed models, and devising methods to automate more and more of the process. Isn't one supposed to gain understanding in science? This tension between understanding and results has

no doubt been around since speech evolved. Should we try to understand how people speak and hear speech, or should we try to make machines that can do useful things, or both? In the 1980s or earlier, in defense of modeling *human* speech processes (an approach already losing ground to more automated approaches) people said, "You can't get to the moon by climbing a tree." That is, measurements showing small progress may not be significant if the task is enormous. Finding a local hill may improve the view, but if you don't look around you might miss the mountain, or the rocket ready to take off. A recent *Science* article (Wilks, 2007) suggests there is still an active discussion concerning machine learning of structure vs. discovering of linguistic structure.

Once any world view takes hold and much is invested in it, it becomes difficult for a competing model to thrive, especially at first. The current statistical models in speech recognition, hidden Markov models, have been around for a few decades and have proved quite useful. However, as Digalakis has pointed out, we know that some of the assumptions in the models are wrong. For example, the assumption that the sound you just made is independent of the sound you will make next is clearly false. Nonetheless much of the current effort in speech modeling focuses on adjusting the boundaries of class identification, rather than on adjusting the models themselves. Statistical pattern matching is a wonderful tool, but its best use is 'ignorance modeling' (see for example, Makhoul and Schwartz, 1986). That is, we should use the statistical models to model things we don't know and try to ensure that we are adequately modeling what we do know. We should not use the success and power of statistics as an excuse to neglect modeling our knowledge, and there are initial efforts in this area. For example, Moore 2007 surveys much of what we know about humans and speech and outlines an interesting approach to taking humans into account in human-machine interactions.

Regardless of the approach, or combination of approaches, the challenges for ASR, as for humans, remain: noise of various sorts, speaker variability, speaking style variability, and increasingly complex tasks. Studying human communication certainly cannot hurt ASR -- human communication is an existence proof of a system that took quite a long development time and does quite well at the task, and what is more defines what the task means. However, machines are not people, so very likely there will remain things ASR is better at and things people are better at, and perhaps our goal should be to understand both so that humans can effectively collaborate with our automated systems, as both evolve.

Predictions. Predictions can be great fun, at least until later when we see how wrong we were. Many predictions are overly-optimistic. Roger K. Moore of the University of Sheffield, UK, has taken a survey every 6 years since 1997. He took some predictions from various pundits added a few at each survey and re-asked attendees at a speech conference when they expected these predictions to come about. In general, people were overly optimistic. In 1997 the median response for when a prediction would come true was 2010. In 2009, the median response to the same set of predictions was 2028 --- the future seems to be getting farther away! This is to be expected since as we learn more we see challenges that were not foreseen. However, in all three sample years, the majority of responses to the prediction of when no more speech research would be needed was "never". Of course there could be some bias in this sample of speech researchers.

In my AAAS talk 10 years ago, I predicted the merging of the properties of speech and text, though I was careful then and am careful now, not to say when that merger might happen because the process is so

gradual. But in fact, speech is becoming more like text in our ability to search speech for content, and in our ability to use speech as a form of input that substitutes for typed input.

Speech and language are inherently **social** constructs, and that is not likely to change. We have many years of evolutionary forces related to our abilities to produce sounds that can be perceived by those with whom we communicate. Thus, the information residing in speech is constrained by what sounds humans can both produce and perceive. It is also constrained by the tension between the needs of the speaker (who might want to expend minimal effort in thought or articulation) and those of the hearer (who might want to expend minimal effort in hearing and understanding). At the very least, ASR researchers should study human speech perception as an example of highly sophisticated speech recognition suited for many of our needs; at best, we should study human speech recognition as a model we can learn from and improve on. Perhaps one day we will have technology that gives speech all the attributes we enjoy in text (persistence over time and place, easy search and retrieval) and more (summarization, inference generation, etc.). Perhaps one day that vision will no longer be considered 'technology' but will seem as natural as speaking and writing do today.

Thanks!

Special thanks for providing papers, thoughts and helpful suggestions to Martin Cooke, Vas Digalakis, Sadaoki Furui, and Roger K. Moore. I also thank many others who generously helped with ideas for this presentation, including: Sondra Ahlen, Fil Aleva, Francoise Beaufays, Joe Campbell, Rolf Carlson, Gerard Chollet, Mike Cohen, Deborah Dahl, Farzad Ehsani, Juan Gilbert, John Makhoul, Bill Meisel, Ariane Nabeth, Joe Picone, Alex Rudnicky, Paul Sawyer, Malcolm Slaney, Michel Stella, Gary Strong, Orith Toledo, Carl Turner, Fuliang Weng, Steve Young. There are many more to thank – all those who helped me think about speech over many years. I thank you all.

References

Apple, 1987, Knowledge Navigator, http://www.digibarn.com/collections/movies/knowledge-navigator.html

Bacchiani, M., Beaufays, F., Schalkwyk, J. , Schuster, M., Strope, B. 2008. "Deploying GOOG-411: Early Lessons in Data, Measurement, and Testing,", Proc. ICASSP.

Cooke, M., Hershey, J. R., and Rennie, S. J. 2010. "Monaural speech separation and recognition challenge", Computer Speech and Language 24 (2010) 1–15.

Digalakis, V. 2010. Personal communication and in presentations.

Makhoul, J. and Schwartz, R. 1986. "Ignorance modeling" in variability and Invariance in Speech Processes., J. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc. pp. 344-345.

Moore, R. K. 2005. Results from a survey of attendees at ASRU 1997 and 2003, INTERSPEECH. Lisbon.

Moore, R. K., 2007, PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction" IEEE Transactions on Computers, 56: 9, pp. 1176 – 1188.

Saturday Night Live, Julie the Operator Lady, 2006, http://www.hulu.com/watch/10409/saturday-night-live-julie-the-operator-lady

Shen, W., Olive, J., Jones, D., 2008. "Two Protocols Comparing Human & Machine Phonetic Discrimination Performance in Conversational Speech", Interspeech 2008, Brisbane, Australia.

Sroka, J. J., and Braida, L. D. 2005. "Human and machine consonant recognition," Speech Commun. 45, 401–423.

Urbina, I. 2004. "Your Train Will Be Late, She Says Cheerily" The New York Times, November 24, 2004
(http://www.nytimes.com/2004/11/24/nyregion/24voice.html?_r=1&ex=1178596800&en=fd4f7bf582e96fcf&ei=5070**)**

Wilks, Y. 2007. "Is there progress on talking sensibly to machines?" Science Vol. 318, Nov'07, pp. 927-928.