Using Speech Technologies for Information Access: Does it Require Getting Involved in Mechanisms of Mind and Intelligence?

John Garofolo
Information Access Division
NIST

Speech is arguably man's oldest and most natural form of communication. Speech and language are also inextricably linked to human thought and intelligence. Therefore, the recognition and understanding of spoken and written language was from the beginning an important component of artificial intelligence research. Initial efforts at speech recognition using classic AI techniques were thought to have failed because of the computational limitations of the time. Yet, even with the major advances that have occurred in computing power over the last decade, successful communication with machines using human language remains elusive.

While it's true that speech-based information systems (primarily telephone call centers) are being deployed on almost a daily basis, these systems are typically highly constrained to a particular limited vocabulary and/or discourse domain. If one tries to ask one of these systems a question that is outside of its narrow scope or if one uses a word that is outside of its vocabulary, one is quickly switched over to a human operator, or worse, sent into telephone oblivion. It's somewhat ironic that the current wealth of speech recognition software/services as well as their limitations are largely due to a mass decision by the research community to abandon traditional AI approaches in the mid-to-late 70s and instead focus on probabilistic methods.

When traditional AI approaches became intractable when applied to language, statistical approaches using variations of Hidden Markov models and neural nets were employed. For a time, during the late 80s and early 90s, there were huge improvements in accuracy. To work effectively, these approaches require large amounts of exemplary training data. So, large amounts of money were devoted in the research community to developing transcribed recordings of hundreds of hours of speech. While larger amounts of training data improved accuracy for particular domains, significant generalizable improvements in the technology were really not occurring. For a time, as progress began to stall, larger and larger training corpora – eventually in the hundreds of hours -- were employed. However, it was soon learned that the problem could not be solved with the amount of data that could be reasonably collected and transcribed and it has been suggested that many thousands of hours of speech would be necessary for significantly improved recognition using these techniques. Unfortunately, some researchers now believe that statistical approaches are inherently limited and that further progress down that path is unlikely. There are clearly at least three important components of recognition of speech by humans that are not addressed by existing approaches: robust acoustic modeling, explicit linguistic knowledge, and world/contextual knowledge. Because of these limitations, recognizers were (and continue to be) limited to the types of vocabulary they could recognize and the acoustic conditions within which they could work with any accuracy. Unfortunately, to date, the speech research community by itself has been unable to address these issues.

Yet progress could not have been gauged at all if the community had not had a way to evaluate recognition performance. Therefore, an important factor that cannot be overlooked in the development of speech recognition technologies was the development of evaluation methodologies and practices. In 1987, with DoD sponsorship, NIST began a series of speech recognition evaluations that continue today. Through the years, the evaluation domains have changed and become more realistic and challenging, but the approach and metrics have remained relatively stable. These evaluations use "canned" recordings of speech and are, therefore, repeatable. In the early years of speech technologies, only demo-based anecdotal so-called evaluations were performed and great claims were made which could not be substantiated. This brought about the "dark days" of speech recognition in the 70s when it was realized that speech recognition (as well as other AI technologies) had been "oversold". The NIST evaluations helped to reverse perceptions about speech recognition and demonstrated measurable progress over time as well as provided direct comparisons between systems. As such, these evaluations have also helped to propel the research into productive directions.

Recently, with NIST's assistance, the Defense Advanced Research Projects Agency (DARPA) has begun a new program to explore new approaches in speech recognition technology. The program has set out to improve performance in two ways: 1) By exploring novel approaches for word recognition; and 2) By creating integrated technologies generating enriched transcripts. This new area, deemed "Rich Transcription", is an effort at generating recognition output that contains a variety of metadata in addition to the words that were spoken. The resulting rich transcripts with speakers indicated, sentences marked, etc. can be rendered into human-readable form and are also more useful for other downstream processing technologies such as retrieval, information extraction, summarization, and translation. Further, the metadata can be "fed back" into the recognizer to improve the word recognition itself. This new program provides a "back door" back into the world of AI by integrating linguistic knowledge into the recognition process and by rethinking that acoustic and language modeling process. Since the recognition process will be tightly integrated with deeper linguistic processing to provide an understanding of the structure and meaning of the words that were spoken, this is an initial step back to speech recognition's AI roots. Other efforts at building speech understanding and discourse capabilities are taking a similar approach.

NIST is widening the scope of its speech recognition evaluations to measure the accuracy of the production of the Rich Transcription metadata as well as the words produced by the recognition systems. It is hoped that this new, broader approach will help to propel the recognition and understanding of human speech by machines to a new level which will fulfill one of the promises of AI.