VIEW FROM AVIOS

By William Meisel

Speech Understanding: The Next Leap in Speech Recognition

Deep learning and the data that feeds it will drive speech tech's commercial use

Early attempts to commercialize speech recognition technology were limited by the state of that technology. One typically had to find an application where hands- or eyes-free control of a system was required, and the actions controlled were restricted to a small vocabulary. One early application that found limited success directed a headset-wearing worker to a particular bin to pick up an order, with the worker saying things like "out of stock" or "next."

More advanced speech recognition technology attempted speech-to-text transcription of continuous speech. A major early market that continues is medical transcription (e.g., a radiologist dictating a report while looking at an image). The technology of transcribing speech to text has improved to the point that the current technology has been reported to at least equal the accuracy of human speech recognition, so the limitation isn't the technology itself. Machine learning using deep neural networks—"deep learning"—has driven the increased accuracy of speech-to-text transcription.

The missing piece that would allow speech recognition to have a big commercial impact was natural language processing.

The commercial reality is that dictating a report with a result that is close to what one might create by typing is an acquired skill. Doctors and lawyers often learn this skill and are major users of speech-to-text software. But the nature of their work is such that the automated transcription almost always needs to be reviewed by the author or a skilled transcriptionist, reducing the cost-effectiveness of the automated solution.

These early applications of speech recognition didn't match the science fiction concept of talking to computers. But after decades of limited use of speech recognition, digital systems, thanks to deep learning, are becoming conversational, which clears the path to widespread commercial use. Digital assistants like Apple's Siri, Amazon's Alexa, and Google Assistant try to deal with every request a user speaks, and these popular personal assistants provide a channel for a company to contact its customers, with tools that can be used to develop company- or application-specific "skills" or "actions" that can be summoned with a voice request.

Humans didn't develop speech for writing—they developed it for communication. For speech to be useful in communication, it must be *understood*. Thus, the missing piece that would allow speech recognition to have a big commercial impact was natural language processing (NLP). In the '70s and '80s, "expert systems" that attempted to answer text questions were an early attempt at NLP, with the emphasis on providing answers rather than understanding questions. An expert in a given area was essentially required to write a complex program with a lot of if-then logic. Adding new information to such a system was similar to modifying complex computer code, with all the issues that such a task involves. The failure of expert systems led to what has been called the "AI Winter," a long period where artificial intelligence was considered an overreach and little research was done.

The revival of NLP was made possible by increased computer processing power and memory. One approach applicable to limited contexts was powered by humans picking key phrases that indicated the intent of an utterance. (For example, "account balance" could indicate a request for the amount of money available in a bank account).

The major breakthrough in NLP, however, came with the increasing amounts of available data that allowed the creation of complex statistical/empirical models and the computer processing power to extract those models from large datasets—in other words, deep learning. Because these methods simply modeled the implications of very large datasets, they didn't require human insights (except indirectly in the labeling of the data).

Empirical techniques aren't new, but today's computing capabilities have crossed a threshold that allows using complex models such as deep neural networks. This breakthrough, along with some methodological improvements in the use of that technology, are allowing increasingly more powerful NLP. And the deployment and use of such NLP systems generates the data needed to improve them, a virtuous cycle that will drive continued advances in performance.

Speech *recognition* is no longer the major goal driving commercial adoption. Speech *understanding* is driving today's high-volume applications. And it's coming just in time to overcome the increasing complexity of today's digital systems and applications. ☒

William Meisel is the president of TMA Associates.