The long quest for technology that understands speech as well as a human

December 4 2015, by Allison Linn

Sitting in his office overlooking downtown Bellevue, Washington,

Microsoft's Fil Alleva is talking about the long and sometimes difficult road he and other speech recognition experts have taken from the early work of the 1970s to the situation he is in today, where he can turn to his computer and say, "Cortana, I want a pizza" and get results.

The conversation quickly drifts deeply into the technology that makes something like that possible, and then Alleva pauses.

"What we all had in the back of our minds, whether we say it or not, was C-3PO," he admits with a grin.

The personable "Star Wars" character who can understand and speak millions of languages may not have been the only inspiration for the world's leading researchers – some also will say that the universal translator that was featured prominently in "Star Trek" spurred their dreams along.

But regardless of whether they were "Star Wars" fans or "Star Trek" loyalists, one thing is clear: The quest to create a computer that can understand spoken language as well as a person was for years so fanciful that the only thing to compare it to was science fiction.

These days, a game console that understands voice commands, apps that can translate your conversation in real time and a virtual assistant that provides you with the numbers of nearby pizza places are all fact, not fiction.

These systems not only exist but are getting better every day, thanks to improvements in data availability, computing power and a subfield of artificial intelligence called machine learning, in which systems improve as they take in more data.

Within just a few years, some researchers now believe, these

technologies could reach a point where computers can understand the words people are saying about as well as another person would.

"We are reaching that inflection point," said Xuedong Huang, the company's chief speech scientist, who, along with Alleva, has been a leading force in speech research and product deployment at Microsoft.

The improvements are already fundamentally changing how we use technology. Maybe we dictate texts instead of typing them or speak to our favorite GPS app instead of thumbing in addresses while speeding down the highway.

Cortana, the Microsoft virtual assistant, now sits front and center on Windows 10, inviting users to "Ask me anything," becoming what Microsoft CEO Satya Nadella has described as a "third runtime." That alone fulfilled a dream Huang has had since he was hired to start Microsoft's speech initiatives in 1993.

'It's just so natural'

Perhaps the biggest marker of speech recognition's success is that we have started to use voice recognition instinctively, without thinking, and with the expectation that they will work for us.

"When machine learning works at its best, you really don't see the effort. It's just so natural. You see the result," says Harry Shum, the executive vice president in charge of Microsoft's Technology and Research group.

It also raises tantalizing possibilities for improving communication and productivity, and just making our lives easier.

Consider this: Until recently, if an English speaker who didn't know a word of Mandarin wanted to talk to someone living in China, they either

had to face the daunting task of learning enough of the language to get by or they had to hire another person to translate the conversation for them.

These days, they can simply have the conversation, in real time, using Skype Translator.

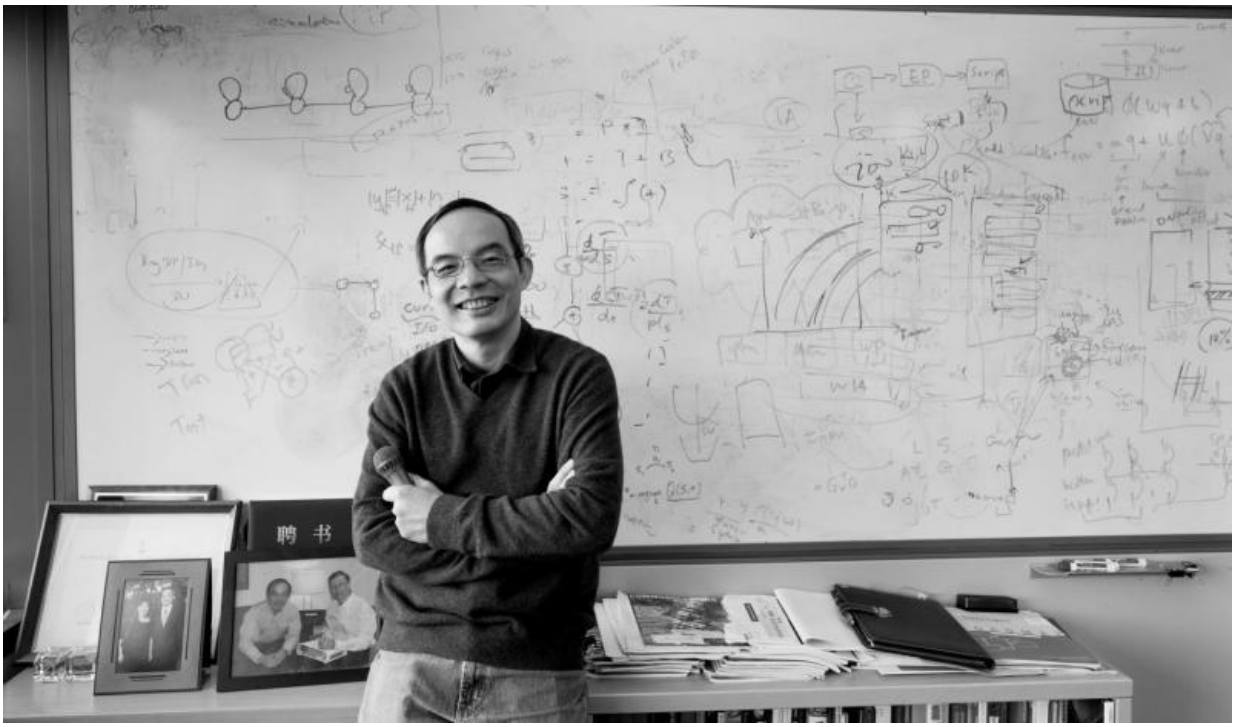
The ability to carry on a conversation in two different languages, and to see the other person's gestures and facial expressions in real time, is one of of several ways in which instant translation is becoming more and more mainstream.

"The language barrier is going to be essentially nonexistent in four years, for the major languages and the major scenarios," said Arul Menezes, who heads the Machine Translation team at Microsoft Research.

It's an amazing idea, and one that language experts have been waiting for, rather impatiently, for years. Menezes started working on the first version of Microsoft Translator while he was on paternity leave with his daughter. She's in high school now.

"We all imagined it. We just didn't know it would take so long to actually work," he said.

A few breakthroughs, and a lot of hard work



Xuedong Huang. Credit: Scott Eklund/Red Box Pictures

Menezes is a member of a fairly large club of researchers and engineers who have been chipping away at speech recognition and translation improvements for decades.

Another is Huang. When he entered graduate school in China in the early 1980s, the first PCs had just been introduced and Chinese users were already seeing a major limitation. The Chinese language itself, with its many characters and variants, just was not conducive to using a traditional keyboard.

Huang – who many people will tell you is an optimist by nature – figured a fix should be easy enough.

"I thought I could solve the problem doing my Ph.D. thesis, by delivering an awesome Chinese dictation engine running on IBM PCs," Huang said.

In some ways Huang is still working on that same graduate school problem, but he is more optimistic than ever that a solution is within reach.

"Speech technology is so close to the level of human performance," he said. "I believe in the next three years we'll reach parity."

Ask Huang, Alleva or Shum why speech recognition models have gotten so much better, and they won't tell you about one big "a-ha" moment. Instead, they'll talk about an improvement here, a breakthrough there and those wonderful times when it all came together thanks to a lot of hard work.

"In research in particular we can take a long-term approach," Shum said. "Research is a marathon."

Data, computing power and machine learning

The basic ingredients for great speech recognition haven't changed much in the decades since Huang, Alleva and Shum met for the first time while studying at Carnegie Mellon University – but they have gotten a lot better.

The first thing you need is data. For a computer to learn to identify sounds, it needs lots and lots of examples to learn from. As more people use tools like Skype Translator or Cortana, those tools are able to get better and better because they have more examples to learn from. Huang calls this influx of usage the oxygen that is fueling speech recognition improvements.

The second ingredient is computing power. Not too long ago, that was limited to whatever a person had on their personal computer or mobile gadget. Now, thanks to cloud computing, there is exponentially more computing power for speech recognition than ever before, even if it's invisible to you.

Finally, you need great machine learning algorithms. Speech experts have used many tools for machine learning over the years, with exotic-sounding names like Gaussian Mixture Models and Hidden Markov Models. A few years ago, Microsoft and other researchers hit on the idea of using a tool called deep neural networks to train computers to better understand speech.

The deep neural networks weren't new, but the way of using them was. The technique worked very well, and now researchers are applying the same technology to other areas of computing, such as computer vision, machine translation, image recognition and automatic image captioning.

Right! Right? Write!

At this point, researchers say, the biggest hurdles in getting a computer to understand speech as well as a human relate to the challenging environments that people encounter.

"In some controlled circumstances, we already have excellent quality speech recognition," said Geoffrey Zweig, who manages Microsoft Research's speech and dialogue group. "In others, we have a long way to go."

Speech recognition tools still don't do very well in noisy, crowded or echo-laden places, and they aren't as good with poor hardware such as low-quality microphones or people talking from far away. They also can struggle when people speak quickly or quietly, or have an accent. It's also

sometimes hard for computers to understand children and elderly speakers.

Microsoft is trying to address that problem with technology such as Microsoft Project Oxford's Custom Recognition Intelligent Services, a forthcoming tool that lets developers build products that deal with those kinds of challenges.

Then there's the much bigger hurdle: Comprehension.

The act of understanding what someone is saying is very different from comprehending the subtle nuances of a person's speech. At the same time as researchers are perfecting language understanding, they also are working on a more nuanced problem: Helping computers understand when a person is enthusiastically shouting "Right!" or sarcastically murmuring "Right?" or brusquely instructing someone to "Write!"

"Normal people, when they think about speech recognition, they want the whole thing," said Hsiao-Wuen Hon, who is managing director of Microsoft Research Asia and also a renowned speech researcher. "They want recognition, they want understanding and they want an action to be taken."

Hon said that means you not only have to solve speech recognition, but you also need to solve natural language, text-to-speech, action planning and execution. Hon refers to this as a system that is "AI complete."

So far at least, humans are much better than computers at understanding these subtle cues. Computer scientists at Microsoft and elsewhere are attacking natural language processing in the same way they went after speech recognition – by gathering data and helping computers learn from it.

Still, it's a bigger challenge in part because rules of natural language understanding are fuzzy. As Hon rightly notes, humans don't always say what they mean. In fact, one could argue that one of the hardest things about being a human is figuring out what other people are trying to communicate to you, and getting them to figure out what you are trying to tell them.

"A lot of times we say that even the people who are closest to us don't understand us," Hon said.

Hearing, but also seeing and even old-fashioned typing

No matter how good we get at creating tools that understand speech, no one expects people to give up the keyboard completely. That's because speech has limitations: It isn't private, for example, and anyone who's ever tried to dictate a paper knows it isn't often conducive to the creative process.

It also isn't exactly reflective of how humans really communicate, not just with their words but also with unspoken cues like facial expressions and gestures.

As speech experts continue to try to solve the problem of speech understanding, they also are thinking more and more about the role speech plays in that bigger goal: To create technology that understands language but also does things like recognize faces and respond to gestures.

"A lot of this is going to depend on building speech within a system," Alleva said. "There's some use for these things in isolation, but there are exponentially more uses when they are coupled together."

Provided by Microsoft