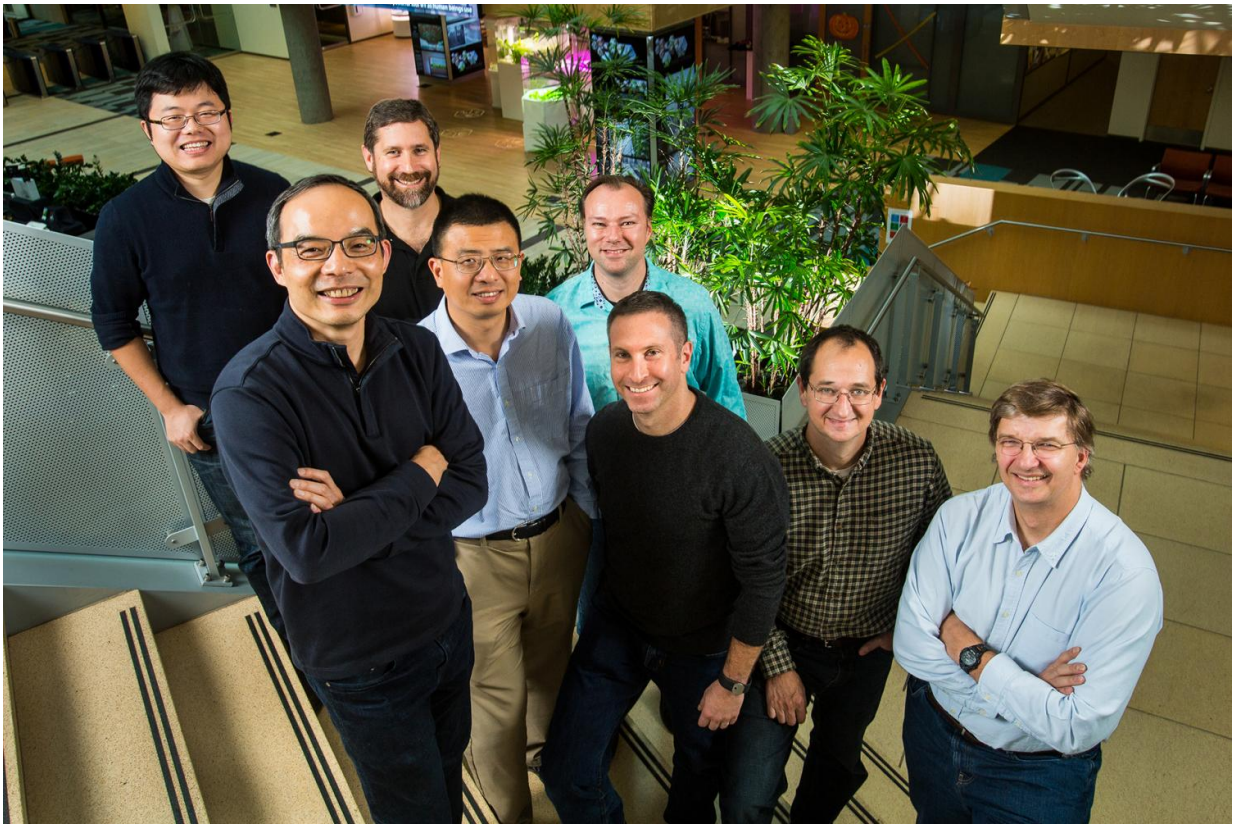
Microsoft claims its new speech recognition system on par with human capabilities

October 25 2016, by Bob Yirka



Microsoft researchers from the Speech & Dialog research group include, from back left, Wayne Xiong, Geoffrey Zweig, Xuedong Huang, Dong Yu, Frank Seide, Mike Seltzer, Jasha Droppo and Andreas Stolcke. (Photo by Dan DeLong)

(Tech Xplore)—Engineers at Microsoft have written a paper describing

their new speech recognition system and claim that the results indicate that their system is as good at recognizing conversational speech as humans. The neural network-based system, the team reports, has achieved a historic achievement—a word rate error of 5.9 percent—making it the first ever below 6 percent, and more importantly, demonstrating that its performance is equal to human performance—they describe it as "human parity." They have uploaded their paper to Cornell's *arXiv* preprint server.

The neural network was taught using recordings made and released by the U.S. National Institute of Standards and Technology—the recordings were created for the purpose of research and included both single-topic and open-topic conversations between two people talking on the telephone. The researchers at Microsoft found that their system had an error rate of 5.9 percent on the single-topic conversations and 11.1 percent on those that were open ended.

As a side note, the researchers report that they also tested the speech recognition skills of humans by having the same phone conversations from NIST sent to a third-party transcription service, which allowed for measuring error rates. They were surprised to find the error rate was higher than expected—5.9 for the single topic conversations and 11.3 percent for open-ended conversations. These findings are in sharp contrast to the general consensus in the scientific community that humans on average have a 4 percent error rate.

The team reports that they believe they can improve their system even more by overcoming obstacles that still confuse their system—namely backchannel communications. These are noises people make during conversation that are not words but still have meaning, such as "uh," "er," and "uh-huh." The neural network still has a hard time figuring out what to do with such noises. We humans use them to allow for pauses, to signify understanding or to communicate uncertainty—or to cue another

speaker, such as to signify they should continue with whatever they were talking about.

The researchers also report that the new technology will be used to improve Microsoft's commercial speech recognition system, known as Cortana, and that work will continue both in improving error rates and in getting their system to better understand what the transcribed words actually mean.

More information: Achieving Human Parity in Conversational Speech Recognition, arXiv:1610.05256 [cs.CL]
arxiv.org/abs/1610.05256

Abstract
Conversational speech recognition has served as a flagship speech recognition task since the release of the DARPA Switchboard corpus in the 1990s. In this paper, we measure the human error rate on the widely used NIST 2000 test set, and find that our latest automated system has reached human parity. The error rate of professional transcriptionists is 5.9% for the Switchboard portion of the data, in which newly acquainted pairs of people discuss an assigned topic, and 11.3% for the CallHome portion where friends and family members have open-ended conversations. In both cases, our automated system establishes a new state-of-the-art, and edges past the human benchmark. This marks the first time that human parity has been reported for conversational speech. The key to our system's performance is the systematic use of convolutional and LSTM neural networks, combined with a novel spatial smoothing method and lattice-free MMI acoustic training.

Microsoft blog: blogs.microsoft.com/next/2016/ … -speech-recognition/