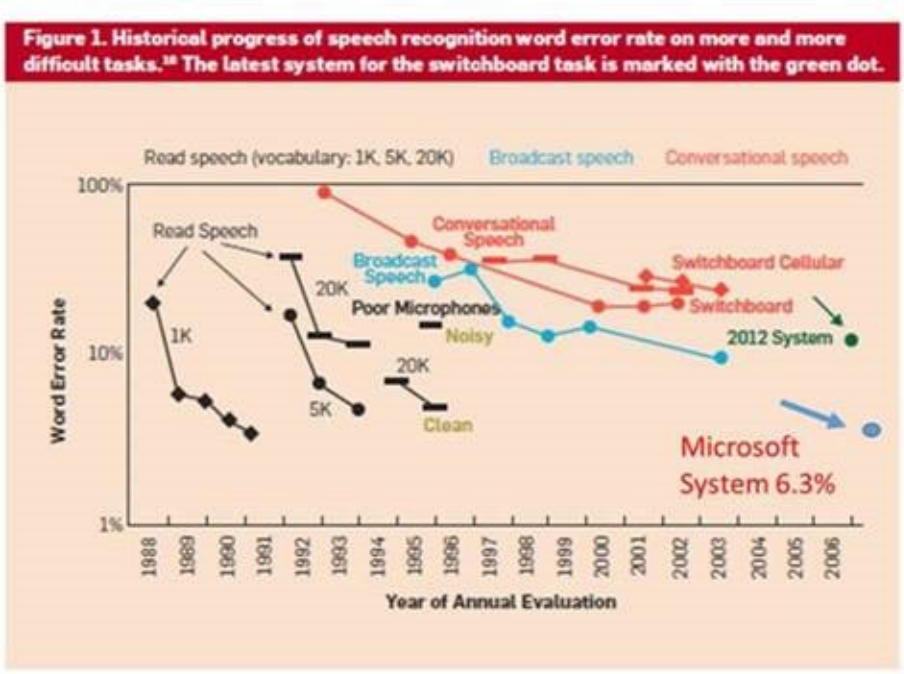
Microsoft researchers in test achieve impressively low error rate for conversational speech recognition system

September 18 2016, by Nancy Owano



Credit: Microsoft

(Tech Xplore)—The languages that we speak: how pervasive will they be in the computing of tomorrow? We are often being told that we are getting closer and closer to computers understanding our words as easily as a human beside us.

Now Microsoft researchers have every reason to feel especially proud. According to reports, Microsoft has stepped in front in the race for supremacy in speech recognition.

The company has claimed a significant test result in their quest for machines to understand speech. The study describing their work has been posted arXiv server. The title is "The Microsoft 2016 Conversational Speech Recognition System." Authors are eight: W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig.

Wall Street Pit had a report about their work, one of a number of sites paying attention to what Microsoft researchers achieved. The Microsoft team turned to "a conversational telephone speech recognition test used as an industry standard," said *Wall Street Pit*. That test is the "US National Institute of Standards and Technology (NIST) 2000 Switchboard speech recognition task."

Chief speech scientist for Microsoft, Xuedong Huang, said their researchers achieved a word error rate (WER) of 6.3%, considered the lowest in the industry.

Richard Eckel posted a piece about it, too, on the Microsoft site. The posting noted some features of their efforts. Earlier this year, Microsoft researchers won a computer vision challenge by using "a deep residual neural net system that utilized a new kind of cross-layer network connection."

It also said that "Another critical component to Microsoft researchers' recent success is the Computational Network Toolkit. CNTK implements sophisticated optimizations that enable deep learning algorithms to run an order of magnitude faster than before. A key step forward was a breakthrough for parallel training on graphics processing

units, or GPUs."

(GPUs are known for computer graphics, but researchers find they are also very good for processing complex algorithms such as the ones used to understand speech, the posting said.)

As for the significance of the error rate, "Last weekend, the international conference speech communication and technology called 'Interspeech' was held in San Francisco," said *Wall Street Pit*. "During the event, IBM proudly announced that it was able to reach a WER of only 6.6%. Over two decades ago, the top error rate of the best published research system for computer speech recognition was at 43%."

The authors stated, "Our best single system achieves an error rate of 6.9% on the NIST 2000 Switchboard set. We believe this is the best performance reported to date for a recognition system not based on system combination. "

Liam Tung in *ZDNet* noted progress in this field. Tung wrote that "20 years ago the lowest error rate in speech recognition was 43 percent and that was achieved by IBM in 1995. By 2004, IBM had cut its error rate to 15.2 percent."

Tung noted that "However, these days with more research funds being funnelled into deep neural networks, tech giants are boasting error rates of well below 10 percent, but not quite at a level that exceeds human-level accuracy, which IBM estimates to be at about four percent."

In describing the system, the authors said, "Inspired by machine learning ensemble techniques, the system uses a range of convolutional and recurrent neural networks."

What distinguishes their work from previous work was explained in the

paper. "Compared to earlier applications of CNNs to speech recognition, our networks are much deeper, and use linear bypass connections across convolutional layers."

Tung remarked that "Like its rivals, Microsoft has made artificial intelligence a key plank in its strategy for human-computer interaction with voice-based platforms such as Cortana set to play a key role in enabling computing in wearables, mobile, the home, vehicles, and the enterprise."

More information: Blog: blogs.microsoft.com/next/2016/ … 22keqx103m2j5pa2aoeg

Paper: The Microsoft 2016 Conversational Speech Recognition System, arXiv:1609.03528 [cs.CL] arxiv.org/abs/1609.03528

Abstract
We describe Microsoft's conversational speech recognition system, in which we combine recent developments in neural-network-based acoustic and language modeling to advance the state of the art on the Switchboard recognition task. Inspired by machine learning ensemble techniques, the system uses a range of convolutional and recurrent neural networks. I-vector modeling and lattice-free MMI training provide significant gains for all acoustic model architectures. Language model rescoring with multiple forward and backward running RNNLMs, and word posterior-based system combination provide a 20% boost. The best single system uses a ResNet architecture acoustic model with RNNLM rescoring, and achieves a word error rate of 6.9% on the NIST 2000 Switchboard task. The combined system has an error rate of 6.3%, representing an improvement over previously reported results on this benchmark task.