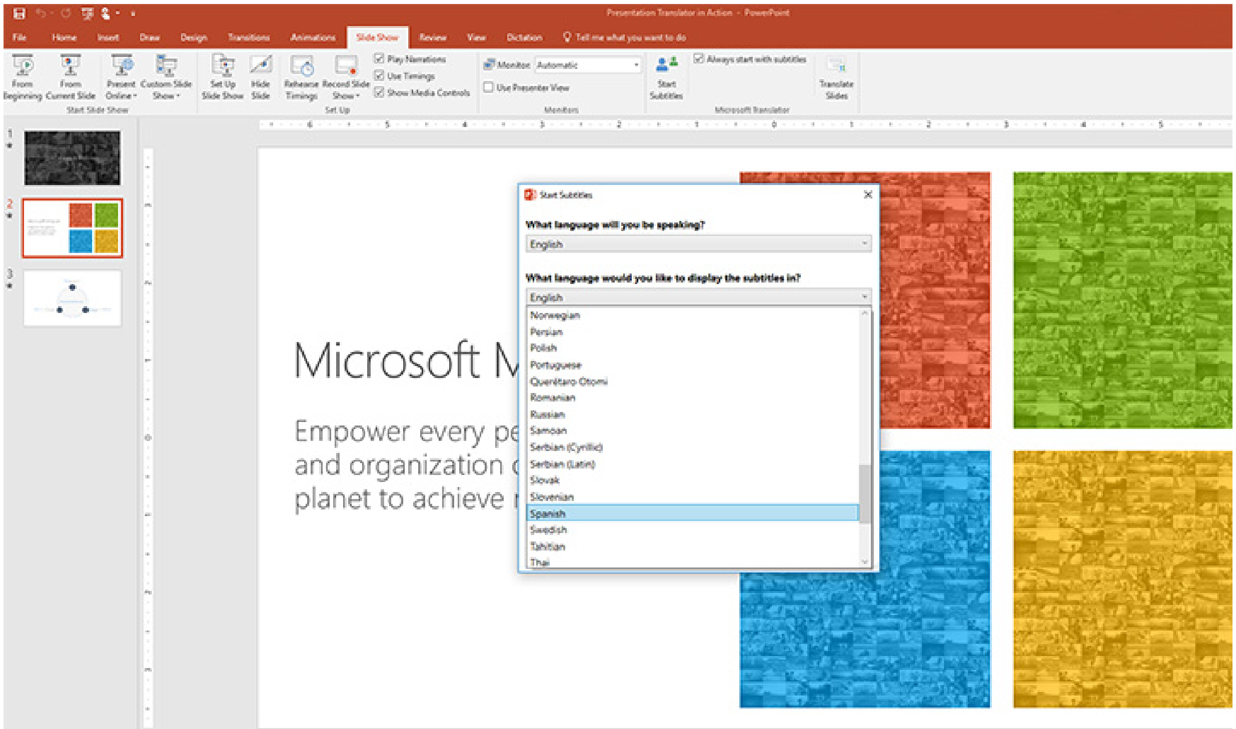
TechXplore

Machine voice recognition reaches human parity

August 21 2017, by Xuedong Huang



Start Subtitles

What language will you be speaking?

English

What language would you like to display the subtitles in?

English

Norwegian
Persian
Polish
Portuguese
Querétaro Otomi
Romanian
Russian
Samoan
Serbian (Cyrillic)
Serbian (Latin)
Slovak
Slovenian
Spanish
Swedish
Tahitian
Thai

Advances in speech recognition have created services such as Speech Translator, which can translate presentations in real-time for multi-lingual audiences.

Last year, Microsoft's speech and dialog research group announced a milestone in reaching human parity on the Switchboard conversational speech recognition task, meaning we had created technology that recognized words in a conversation as well as professional human transcribers.

After our transcription system reached the 5.9 percent word error rate that we had measured for humans, other researchers conducted their own study, employing a more involved multi-transcriber process, which yielded a 5.1 human parity word error rate. This was consistent with prior research that showed that humans achieve higher levels of agreement on the precise words spoken as they expend more care and effort. Today, I'm excited to announce that our research team reached that 5.1 percent error rate with our speech recognition system, a new industry milestone, substantially surpassing the accuracy we achieved last year. A technical report published this weekend documents the details of our system.

Switchboard is a corpus of recorded telephone conversations that the speech research community has used for more than 20 years to benchmark speech recognition systems. The task involves transcribing conversations between strangers discussing topics such as sports and politics.

We reduced our error rate by about 12 percent compared to last year's accuracy level, using a series of improvements to our neural net-based acoustic and language models. We introduced an additional CNN-BLSTM (convolutional neural network combined with bidirectional long-short-term memory) model for improved acoustic modeling. Additionally, our approach to combine predictions from multiple acoustic models now does so at both the frame/senone and word levels.

Moreover, we strengthened the recognizer's language model by using the entire history of a dialog session to predict what is likely to come next, effectively allowing the model to adapt to the topic and local context of a conversation.

Our team also has benefited greatly from using the most scalable deep learning software available, Microsoft Cognitive Toolkit 2.1 (CNTK), for exploring model architectures and optimizing the hyper-parameters of our models. Additionally, Microsoft's investment in cloud compute infrastructure, specifically Azure GPUs, helped to improve the effectiveness and speed by which we could train our models and test new ideas.

Reaching human parity with an accuracy on par with humans has been a research goal for the last 25 years. Microsoft's willingness to invest in long-term research is now paying dividends for our customers in products and services such as Cortana, Presentation Translator, and Microsoft Cognitive Services. It's deeply gratifying to our research teams to see our work used by millions of people each day.

Many research groups in industry and academia are doing great work in speech recognition, and our own work has greatly benefitted from the community's overall progress. While achieving a 5.1 percent word error rate on the Switchboard speech recognition task is a significant achievement, the speech research community still has many challenges to address, such as achieving human levels of recognition in noisy environments with distant microphones, in recognizing accented speech, or speaking styles and languages for which only limited training data is available. Moreover, we have much work to do in teaching computers not just to transcribe the words spoken, but also to understand their meaning and intent. Moving from recognizing to understanding speech is the next major frontier for speech technology.

More information: Technical Report: The Microsoft 2017 Conversational Speech Recognition System: www.microsoft.com/en-us/resear … -recognition-system/

Provided by Microsoft

Citation: Machine voice recognition reaches human parity (2017, August 21) retrieved 10 April 2024 from https://techxplore.com/news/2017-08-machine-voice-recognition-human-parity.html