Speech recognition leaps forward

29 August 2011, By Janie Chang

During Interspeech 2011, the 12th annual Conference of the International Speech Communication Association being held in Florence, Italy, from Aug. 28 to 31, researchers from Microsoft Research will present work that dramatically improves the potential of real-time, speaker-independent, automatic speech recognition.

Dong Yu, researcher at Microsoft Research Redmond, and Frank Seide, senior researcher and research manager with Microsoft Research Asia, have been spearheading this work, and their teams have collaborated on what has developed into a research breakthrough in the use of artificial neural networks for large-vocabulary speech recognition.

The Holy Grail of Speech Recognition

Commercially available speech-recognition technology is behind applications such as voice-to-text software and automated phone services. Accuracy is paramount, and voice-to-text typically achieves this by having the user "train" the software during setup and by adapting more closely to the user's speech patterns over time. Automated voice services that interact with multiple speakers do not allow for speaker training because they must be usable instantly by any user. To cope with the lower accuracy, they either handle only a small vocabulary or strongly restrict the words or patterns that users can say.

The ultimate goal of automatic speech recognition is to deliver out-of-the-box, speaker-independent speech-recognition services-a system that does not require user training to perform well for all users under all conditions.

"This goal has increased importance in a mobile world," Yu says, "where voice is an essential interface mode for smartphones and other mobile devices. Although personal mobile devices would be ideal for learning their user's voices, users will continue to use speech only if the initial experience, which is before the user-specific models can even be built, is good."

Speaker-independent speech recognition also addresses other scenarios where it's not possible to adapt a speech-recognition system to individual speakers-call centers, for example, where callers are unknown and speak only for a few seconds, or web services for speech-to-speech translation, where users would have privacy concerns over stored speech samples.

Renewed Interest in Neural Networks

Artificial neural networks (ANNs), mathematical models of the low-level circuits in the human brain, have been a familiar concept since the 1950s. The notion of using ANNs to improve speech-recognition performance has been around since the 1980s, and a model known as the ANN-Hidden Markov Model (ANN-HMM) showed promise for large-vocabulary speech recognition. Why then, are commercial speech-recognition solutions not using ANNs?

"It all came down to performance," Yu explains. "After the invention of discriminative training, which refines the model and improves accuracy, the conventional, context-dependent Gaussian mixture model HMMs (CD-GMM-HMMs) outperformed ANN models when it came to large-vocabulary speech recognition."

Yu and members of the Speech group at Microsoft Research Redmond became interested in ANNs when recent progress in building more complex "deep" neural networks (DNNs) began to show promise at achieving state-of-the-art performance for automatic speech-recognition tasks. In June 2010, intern George Dahl, from the University of Toronto, joined the team, and researchers began investigating how DNNs could be used to improve large-vocabulary speech recognition.

"George brought a lot of insight on how DNNs work," Yu says, "as well as strong experience in

training DNNs, which is one of the key components in this system."

A speech recognizer is essentially a model of fragments of sounds of speech. An example of such sounds are "phonemes," the roughly 30 or so pronunciation symbols used in a dictionary. State-of-the-art speech recognizers use shorter fragments, numbering in the thousands, called "senones."

Earlier work on DNNs had used phonemes. The research took a leap forward when Yu, after discussions with principal researcher Li Deng and Alex Acero, principal researcher and manager of the Speech group, proposed modeling the thousands of senones, much smaller acoustic-model building blocks, directly with DNNs. The resulting paper, Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition by Dahl, Yu, Deng, and Acero, describes the first hybrid context-dependent DNN-HMM (CD-DNN-HMM) model applied successfully to large-vocabulary speech-recognition problems.

"Others have tried context-dependent ANN models," Yu observes, "using different architectural approaches that did not perform as well. It was an amazing moment when we suddenly saw a big jump in accuracy when working on voice-based Internet search. We realized that by modeling senones directly using DNNs, we had managed to outperform state-of-the-art conventional CD-GMM-HMM large-vocabulary speech-recognition systems by a relative error reduction of more than 16 percent. This is extremely significant when you consider that speech recognition has been an active research area for more than five decades."

The team also accelerated the experiments by using general-purpose graphics-processing units to train and decode speech. The computation for neural networks is similar in structure to 3-D graphics as used in popular computer games, and modern graphics cards can process almost 500 such computations simultaneously. Harnessing this computational power for neural networks contributed to the feasibility of the architectural model.

In October 2010, when Yu presented the paper to an internal Microsoft Research Asia audience, he spoke about the challenges of scalability and finding ways to parallelize training as the next steps toward developing a more powerful acoustic model for large-vocabulary speech recognition. Seide was excited by the research and joined the project, bringing with him experience in large-vocabulary speech recognition, system development, and benchmark setups.

Benchmarking on a Neural Network

"It has been commonly assumed that hundreds or thousands of senones were just too many to be accurately modeled or trained in a neural network," Seide says. "Yet Yu and his colleagues proved that doing so is not only feasible, but works very well with notably improved accuracy. Now, it was time to show that the exact same CD-DNN-HMM could be scaled up effectively in terms of training-data size."

The new project applied CD-DNN-HMM models to speech-to-text transcription and was tested against Switchboard, a highly challenging phone-call transcription benchmark recognized by the speech-recognition research community.

First, the team had to migrate the DNN training tool to support a larger training data set. Then, with help from Gang Li, research software-development engineer at Microsoft Research Asia, they applied the new model and tool to the Switchboard benchmark with more than 300 hours of speech-training data. To support that much data, the researchers built giant ANNs, one of which contains more than 66 million inter-neural connections, the largest ever created for speech recognition.

The subsequent benchmarks achieved an astonishing word-error rate of 18.5 percent, a 33-percent relative improvement compared with results obtained by a state-of-the-art conventional system.

"When we began running the Switchboard benchmark," Seide recalls, "we were hoping to achieve results similar to those observed in the voice-search task, between 16- and 20-percent relative gains. The training process, which takes about 20 days of computation, emits a new, slightly

more refined model every few hours. I impatiently tested the latest model every few hours. You can't imagine the excitement when it went way beyond the expected 20 percent, kept getting better and better, and finally settled at a gain of more than 30 percent. Historically, there have been very few individual technologies in speech recognition that have led to improvements of this magnitude."

The resulting paper, titled Conversational Speech Transcription Using Context-Dependent Deep Neural Networks by Seide, Li, and Yu, is scheduled for presentation on Aug. 29. The work already has attracted considerable attention from the research community, and the team hopes that taking the paper to the conference will ignite a new line of research that will help advance the state of the art for DNNs in large-vocabulary speech recognition.

Bringing the Future Closer

With a novel way of using artificial neural networks for speaker-independent speech recognition, and with results a third more accurate than what conventional systems can deliver, Yu, Seide, and their teams have brought fluent speech-to-speech applications much closer to reality. This innovation simplifies speech processing and delivers high accuracy in real time for large-vocabulary speech-recognition tasks.

"This work is still in the research stages, with more challenges ahead, most notably scalability when dealing with tens of thousands of hours of training data. Our results represent just a beginning to exciting future developments in this field," Seide says. "Our goal is to open possibilities for new and fluent voice-based services that were impossible before. We believe this research will be used for services that change how we work and live. Imagine applications such as live speech-to-speech translation of natural, fluent conversations, audio indexing, or conversational, natural language interactions with computers."

Provided by Microsoft Corporation