(BLDC) motor commutation, sensor signal sampling, and communication with the central limb controller (LC) via a controller area network (CAN) bus.

The LMC was also designed to monitor local joint temperature, torque, position, current, and rotor position sensors for motor commutation. Custom schematic design and multilayer board fabrication allowed direct integration within the drive module. This approach allowed the drives to be designed as single integrated motor and controller packages, helping to shrink the overall mechanical profile and maximizing performance. Each LMC uses an advanced reduced instruction set computer (ARM)-based processor.

"It really comes down to a matter of functionality," McLoughlin says. "If you think about things like turning a doorknob, that's very difficult to do with a prosthesis." The researchers wanted to create a prosthesis capable of extremely fine dexterity and precision allowing users perform tasks ranging from the mundane, such as turning a doorknob up to and including playing a piano. "We want people to be able to do the very complex things with their fingers that those of us who have an arm and hand do naturally," McLoughlin says.

Studying brain signals is crucial to the team's research, since such data is essential for enabling natural control of the artificial limb. "If you think about moving your arm, or opening and closing your hand, there are areas of your brain that will become active," McLoughlin says. "You can actually see those areas of activity if you do an MRI."

The team searches for specific types of signals in different parts of the brain. "The signal processing typically involves things like pattern recognition, in which we look for patterns of neural activity," McLoughlin says. "We can then begin to use pattern recognition techniques to interpret what the user's intent was."

The researchers need to work quickly. "It's all done in real time," he says. "You filter it, process it, pull out the information, and then convert that data into motor commands."

McLoughlin says he's always amazed by the brain's flexibility and adaptability. "We're looking for very specific structures in the signal, and we come up with a very specific model of the signals we're looking at," he says. "I think it's going to open up a whole new realm of possibilities for assistive devices, particularly for the elderly or people with mobility problems who will be able to use machines in ways that are well beyond what we can do now."

One of the challenges facing the team is finding a way of driving down the sophisticated limb's cost. "Right now this is a research tool," McLoughlin says. "We've had ten or 12 different patients utilize the limb with great success, and we have to look at getting it down to a cost point where it's affordable," he says.

AUTHOR

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona, area.

Ron Schneiderman

Accuracy, Apps Advance Speech Recognition

Technical breakthroughs in speech recognition have been hard to come by, but the technology continues to improve in accuracy and natural language understanding and find its way into a broad range of enterprise and commercial platforms that include health care, e-commerce, telecommunications, and other vertical markets.

In this third in a series of Q&A interviews for *IEEE Signal Processing Magazine (SPM)*, we talked to Li Deng, the principal researcher and research manager of the Deep Learning Technology Center at Microsoft Research, and Vlad Sejnoha, the chief technology officer at Nuance Communications, about current activities and future developments in speech recognition, text-to-speech (TTS), speech-to-speech translation, and related applications.

***IEEE SPM**: Accuracy has been an issue in speech technology since its emergence out of Bell Labs in the 1950s. How has it improved?*

Vlad Sejnoha: We actually have been improving it quite quickly over the years. We have improved the error rate by a consistent amount and there seems to be no end to this. Each year, it's through a different combination of new algorithms, more data, more computation. Different mixes of that. So, when we talk about speech recognition accuracy, it really is a moving target. I think we have passed a magical threshold of usability that means that you can pick up a device today and speak to it and expect to be understood.

In recent years, the focus has been on deep learning. That is the sort of algorithmic underpinning of why we are continuing to improve. In a few years, it might be something else. So, it's a very rapid and dynamically evolving process.

Li Deng: Progress was relatively slow from 1989 to 2009 compared with the last few years after deep learning made inroads into speech recognition. The introduction of deep learning has been one of the major breakthroughs, mostly in the form of deep neural networks. But

Digital Object Identifier 10.1109/MSP.2014.2358717
Date of publication: 5 December 2014

there's expected to be work that may go beyond neural networks.

The concept of deep learning applies in terms of being able to absorb more data and you can actually impart relevant domain knowledge of speech into the system in a hierarchical manner. All of this new technology, especially from the machine-learning community that is heavily overlapping with our speech/language and signal processing communities, has had a tremendous impact in our field. That's why we're quite hopeful that the accuracy will keep improving, even after neural networks reach their potential limit sometime in the future.

Deep learning started its success drastically raising speech recognition accuracy with deep neural networks [around 2010, soon after Microsoft researchers collaborated with Prof. Geoff Hinton in Redmond, Washington]. In the future, I expect that deep neural network will integrate with other forms of deep models in elegant and theoretically appealing ways to achieve better success, with the new capability of not only absorbing big data but more importantly "big knowledge" that would include semantic knowledge in the top-down fashion.

Sejnoha: One thing that should be pointed out, I think, is the importance of good signal acquisition. If you have the best speech recognition system in the world, if you're using it in a very difficult environment, with a mobile device with a microphone that might be moving and not properly positioned to you, it will not do very well. So, we're finding that these fundamental modeling improvements, like deep learning, are being coupled with more and more sophisticated ways of capturing a signal, and that includes things like using multiple microphones that can be configured into a steerable beam that can track the speaker, in combination with voice biometrics where you actually know who is talking and you can tease out the desired signal from interfering signals. And you can use this with other sensors like audio-visual recognition where you can actually use camera feeds to lock in on the speaker. That's becoming extremely important, but improvements there are equal to multiple generations of modeling improvements.

***IEEE SPM**: What is deep learning?*

Sejnoha: It's an approach that tries to model or make decisions about the nature of input signals by organizing layers of very simple processing units, which are very loosely modeled after our understanding of how neurons work. Each processing unit has a number of inputs, and those inputs are weighted. They come from the outputs of other neurons. These neurons are processing nodes. Some of these are weighted inputs and then pass them through in non-linear function that basically says on or off, or some degree of that, and then passes that on to higher layers.

It can model very complex decision spaces. In the past, these were very difficult to train and got stuck in what we call *local optima*. There have been a number of breakthroughs in the training in recent years that help these layered networks reach more global or overall optima. It's called *compositionality*; that you can have relatively few processing units and they can explain very high dimensional spaces. Ever since they became more trainable, this has been applied to a wide variety of problems—acoustic modeling, language modeling, and assigning meaning to patterns.

Interestingly, you can use them to concurrently learn or optimize multiple parts of the recognition process. Traditionally, we would build a so-called feature extractor that takes the audio signal and maps it usually into spectral space that we think is more amenable to further processing. It turns out that these deep neural nets can simultaneously learn a more optimal set of features that also make decisions on the results. And that's exciting, but I want to caution about thinking that the world going forward is all going to be about machine learning, or learning from examples and patterns.

A lot of current cutting-edge work is about how to combine machine learning with techniques that encapsulate our existing knowledge of the world, either through rules or grammar, or explicit knowledge bases where you describe object, concept, and relationships.

***IEEE SPM**: What's the role of artificial intelligence (AI) in improving the accuracy of speech recognition? Is it advancing the technology?*

Sejnoha: AI is a big term, but some aspects of AI are already manifested in today's virtual assistants. A simple example would be a user asking for a restaurant that serves good spaghetti, and it turns out that the back-end services to which you're connected don't understand that because they only understand cuisine types. You would actually take this input user request, and, using explicit reasoning on a knowledge base that relates specific dishes to specific cuisines, it would learn that spaghetti is a form of Italian cuisine. But that's a trivial example. There are a lot more sophisticated ways of relating an input concept language with a backup concept language and deciding what makes sense and doesn't make sense.

But there could be a lot of subtasks in a request that a virtual assistant should be able to do, but there are lots of contingencies. Being able to specify tasks by expressing goals at a high level with automatic back-off strategies versus prescripting every possible interaction is really important in AI.

***IEEE SPM**: To what extent is digital signal processing (DSP) playing a role in the development of speech technology?*

Deng: The community, including many speech recognition, understanding, and machine-learning researchers, are part of our [IEEE Signal Processing] Society, so we are actually thinking about changing the Society's name to the IEEE Signal and Information Processing Society. This would better describe the recent activities of this community [of the Society's members]. Many big companies, and I am thinking about Microsoft, Apple, Google, Baidu, etc., have researchers and engineers working on speech technology problems much more complex than the DSP topics you would see in Oppenheimer's book. We have moved well beyond traditional DSP.

As for deep learning, the people who brought that into large-scale speech technology applications are mainly from our [IEEE] Signal Processing Society. I recently gave a long lecture at the International Conference on Machine Learning, and I delivered the keynote at the

Interspeech Conference held in Singapore in mid-September, reflecting on part of that history with emphasis on the industry–academic collaboration and on what future direction that history points to. A lot of DSP techniques, such as short-time Fourier transforms, cepstral analysis, and linear prediction, which used to be standard analysis as front ends for speech recognition systems, are now becoming mere initialization of low layers of the full deep-learning system, subject to much more important step of end-to-end learning by what is known as back propagation beyond DSP.

However, one important concept of DSP, convolution, has been playing a crucial role in modern deep learning systems. They are called *deep convolutional neural networks*, very popular in image recognition and recently also gaining popularity in speech recognition.

IEEE SPM: *How much progress has been made in implementing TTS, especially in making it more natural?*

Sejnoha: That's a topic that is near and dear to us. The TTS, or speech synthesis field, has moved through some phases through the years. It started out by building what we call some model-based approaches, some mathematical models that are very compact, and the use of algorithms to express how particular text should be mapped–ultimately into a waveform. But they sounded very robotic.

There was a breakthrough in the late 1990s when the first commercial systems started coming out. People discovered that by extracting snippets of real speech and forming them into structured data bases you could, on the fly, concatenate, or glue together, the appropriate segments and they sounded far more natural. The problem with those systems is that they were big. There were lots of different segments that were needed, and it was very difficult to manipulate pitch, duration, and loudness, but it really carried a lot of emotion and naturalness in human speech.

So a lot of the work now is how to build hybrid systems, just like hybrid systems in understanding and combine machine learning and some explicit knowledge. Building TTS that have the quality of the concatenated system, but the ability to manipulate pitch and loudness and volume—all of the prosodic signals based on understanding of the text. So, what we do is take text and apply natural understanding to what is being said and what it means, and use a combination of prerecorded segments and models to try to generate expressive speech synthesis. Deep learning plays a role there as well.

Deng: Deep learning has also been making an inroad into speech synthesis or TTS research since last year. At ICASSP 2013, there were four nice papers on this topic, from different angles and for different aspects of the synthesis problem. They demonstrated more natural subjective speech sound's quality produced by deep learning systems than the previous state-of-the-art, Gaussian-HMM-based statistical methods. More research papers have come out since then. In a sense, it is very intuitive to adopt an original, generative version of deep learning approaches, called the deep believe network, which is quite different from the deep neural network, to deal with speech generation or synthesis problems.

IEEE SPM: *Are there particular challenges at this point in deploying speech technology globally given the need to support many languages and with a high degree of accuracy?*

Deng: You want to have voice systems perform well in noisy environments. These include the conditions where the voice intended to be recognized are mixed with other speakers' voices, such as when playing Xbox or Kinect games with voice control. As was demoed in May this year, Skype translator will be able to perform real-time speech-to-speech translation. Under the conditions where there is no close-talking microphone, noise robustness, especially the robustness against other speakers' voice, in speech recognition component of the system is very important. Human listeners can use attention to focus on the intended speaker, but so far computer systems cannot simulate such ability easily. Deep learning is moving toward solving such difficult problems, with preliminary promising results already seen in the literature.

Before the rise of deep learning, multilingual speech recognition was very difficult in economic terms due to the need to collect data and design dictionaries from many languages. Deep neural networks have drastically reduced this challenge, thanks to the "transfer learning" capability where the upper hidden layers in the deep networks are shown to represent more abstract acoustic features universal across different language. This capability is made possible because acoustic properties of speech, no matter which language it belongs, are shared across languages since they are all generated by the highly constrained human vocal tract, plus the rest of the speech production system. Only deep learning systems can effectively take advantage of such constraints, not the previous systems without hierarchical feature representations.

For many speech recognition applications that are linked closely to downstream processing, semantic understanding of the recognition output and of the end tasks and the final actions taken by the overall system are the final goal. One particular technical challenge here is how to effectively represent semantics and the backend application-domain knowledge. Recent advances in deep learning for natural language processing have provided a very interesting approach where any semantic linguistic entity and simple relation in the knowledge source can be mapped into a continuous-valued vector, called *embedding*. Embedding has been shown to be quite effective for a word, a phrase, a sentence, a paragraph, or even a whole document. These embedded linguistic units can also be used to represent the output of a speech recognizer. Thus, the designs of downstream text processing and speech recognition systems are intimately connected and can be jointly optimized.

Despite such progress, however, semantic representations for more advanced tasks that would require structured representations and complex relations may not be adequately accomplished with vector

(continued on page 125)

Mast and Denise Frauendorfer), Idiap (Dairazalia Sanchez-Cortes, Oya Aran, Laurent Nguyen, Alvaro Marcos, Dinesh Babu Jayagopi, and Jean-Marc Odobez), and other institutions (Tanzeem Choudhury, Cornell University; Marta Marron, University of Alcala, Spain; and Daniel Pizarro, University of Auvergne, France.) I thank all of them, and acknowledge the support by the Swiss National Science Foundation (SONVB and UBImpressed projects) and the European Commission (NOVICOM project).

AUTHOR

Daniel Gatica-Perez (gatica@idiap.ch) is the head of the Social Computing Group at Idiap Research Institute and Maître d'Enseignement et de Recherche at the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland.

REFERENCES

[1] (2014, June 18). American time use survey summary. [Online]. Available: http://www.bls.gov/news.release/atus.nr0.htm

[2] M. Remland, "Uses and consequences of nonverbal communication in the context of organizational life," in *The SAGE Handbook of Nonverbal Communication*, V. Manusov and M. Patterson, Eds. Thousand Oaks, CA: Sage Publications, 2006, pp. 501–521.

[3] J. A. Hall, E. J. Coats, and L. Smith, "Nonverbal behavior and the vertical dimension of social relations: A meta-analysis," *Psychol. Bull.*, vol. 131, no. 6, pp. 898–924, 2005.

[4] R. T. Stein, "Identifying emergent leaders from verbal and nonverbal communications," *Pers. Social Psychol.*, vol. 32, no. 1, pp. 125–135, 1975.

[5] D. Sanchez Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, nos. 2–3, pp. 816–832, June 2012.

[6] D. Sanchez Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez, "Emergent leaders through looking and speaking: From audio-visual data to multimodal recognition," *J. Multimodal User Interfaces* (Special Issue on Multimodal Corpora), vol. 7, nos. 1–2, pp. 39–53, Mar. 2013.

[7] A. S. Imada and M. D. Hakel, "Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews," *Appl. Psychol.*, vol. 62, no. 3, pp. 295–300, 1977.

[8] R. J. Forbes and P. R. Jackson, "Non-verbal behaviour and the outcome of selection interviews," *Occupational Psychol.*, vol. 53, no. 1, pp. 65–72, 1980.

[9] L. S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1018–1031, June 2014.

[10] L. S. Nguyen, J.-M. Odobez, and D. Gatica-Perez, "Using self-context for multimodal detection of head nods in face-to-face interaction," in *Proc. ACM Int. Conf. Multimodal Interaction (ICMI)*, Santa Monica, Oct. 2012, pp. 289–292.

[11] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. S. Nguyen, and D. Gatica-Perez, "Body communicative cue extraction for conversational analysis," in *Proc. IEEE Int. Conf. Face and Gesture Recognition (FG)*, Shanghai, Apr. 2013, pp. 1–8.

[12] D. Frauendorfer, M. Schmid Mast, L. S. Nguyen, and D. Gatica-Perez, "Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example," *J. Nonverb. Behav.* (Special Issue on Contemporary Perspectives in Nonverbal Research), vol. 38, no. 2, pp. 231–245, June 2014.

[13] M. E. Hoque and R. W. Picard, "Rich nonverbal sensing to enable new possibilities in social skills training," *IEEE Comput.*, vol. 47, no. 4, pp. 28–35, Apr. 2014.

[SP]

[special REPORTS] (continued from page 14)

embeddings. Then, the challenge is how to effectively embed the full structure in the appropriate semantic space. If this is done well, the speech recognition component of the overall system will have powerful constraints to exploit, leading to the reduction of its language model's perplexity and improvement of its recognition accuracy.

Sejnoha: I think that the signal acquisition, making sense out of a very noisy world, is a very important challenge and something we have to continue working on. The fundamental modeling and modeling language—I think we're making good progress in these areas. When it comes to extraction and knowing what to do, that borders on AI. How do you define the goal of an interaction with a user in a way that it is efficient and where unexpected intelligent things happen? I think that's still a fairly novel area. You will see a lot of progress there.

The big challenge is connecting to the myriad of forms of content and services that people want to interact with, and part of that is an engineering issue and part of it is the fundamental problem of the promise of the semantic web. We have lots of stuff out there, but it is siloed, it's opaque. It doesn't advertise its capabilities, or describe its knowledge in machine understandable terms. As we get closer to the real Internet of Things, we will do better on that front. When you tell your virtual assistant to turn down your thermostat, they can talk to each other.

IEEE SPM: *What qualifications would be needed for engineers interested in specializing in speech technology? What skill sets would be most helpful?*

Sejnoha: The field has huge multidisciplinary demands. Some background in digital signal processing and modeling is important. Of course, AI and machine learning. Also, software development. And linguistics.

RESEARCHERS INTERVIEWED

Li Deng is the principal researcher and manager of research of the Deep Learning Technology Center at Microsoft Research.

Vlad Sejnoha is the chief technology officer of Nuance Communications.

Editor's Note: This interview was conducted by Ron Schneiderman, a regular contributor to *IEEE SPM*.

[SP]