Introduction to the Special Section on Deep Learning for Speech and Language Processing

IN the past two decades, most work in speech and language processing has used "shallow" models that lack multiple layers of adaptive nonlinear features. Current speech recognition systems, for example, typically use Gaussian mixture models (GMMs), to estimate the observation (or emission) probabilities of hidden Markov models (HMMs), and GMMs are generative models that have only one layer of latent variables. Instead of developing more powerful models, most of the research effort has gone into finding better ways of estimating the GMM parameters so that error rates are decreased or the margin between different classes is increased. The same observation holds for natural language processing (NLP) in which maximum entropy (MaxEnt) models and conditional random fields (CRFs) have been popular for the last decade. Both of these approaches use shallow models whose success largely depends on the use of carefully handcrafted features.

Shallow models have been effective in solving many simple or well-constrained problems, but their limited modeling power can cause difficulties when dealing with more complicated real-world applications. For example, a state-of-the-art GMM-HMM based speech recognition system that achieves less than 5% word error rate (WER) on read English may exceed 15% WER on spontaneous speech collected under real usage scenarios due to variations in environment, accent, speed, co-articulation, and channel.

The discovery of more effective ways of learning multiple layers of features in deep neural networks has created renewed interest in this type of model in the machine learning community, and this motivated us to organize this special section to examine the current potential of using deep models in solving speech and language processing problems.

For the sake of the special section, we defined deep learning techniques as machine learning methods that involve at least three, adaptive nonlinear processing steps from the input to the output. Deep models that learn many layers of features can potentially extract much better information from the input signal for the task in hand (e.g., classification or synthesis), through many layers of nonlinear evidence combination.

We want to emphasize that deep models are not new. They have been in existence for decades. For example, hierarchical (or stacked) HMMs or CRFs and multi-level detection-based systems both are deep models. Even conventional GMM-HMM systems, when combined with several layers of nonlinear or piecewise-linear feature transformation techniques (e.g., the tandem architecture), can be considered to be deep models. However, these existing deep models are quite limited in exploiting the full potential deep learning techniques can bring to advance the state of the art in speech and language processing.

The review process resulted in five accepted papers for publication in this special section. "Deep and Wide: Multiple Layers in Automatic Speech Recognition" by Morgan discusses some of the existing deep models in speech recognition and argues that, in developing more powerful models, increasing the width of each layer of features is at least as important as increasing the depth. The papers "Acoustic Modeling Using Deep Belief Networks" by Mohamed, Dahl, and Hinton, "Sparse Multilayer Perceptron for Phoneme Recognition" by Sivaram and Hermansky, and "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition" by Dahl, Yu, Deng, and Acero all exploit a deep neural network/HMM hybrid architecture for speech recognition. The optimal number of hidden layers in some of the systems described can be as high as eight. While the first two papers focus on recognition of context-independent monophones, the third paper extends the technique to context-dependent triphones and large-vocabulary continuous speech recognition. "Bayesian Sensing Hidden Markov Models" by Saon and Chien discusses a new type of hidden Markov model in which a set of hidden basis vectors and associated weights and precision matrices are jointly optimized.

We hope that this special section will stimulate interest in developing more powerful deep learning techniques for speech and language processing, and we look forward to seeing an increasing amount of high-quality research in this area that improves the state-of-the-art in speech and language processing.

We would like to express our gratitude to the authors of the papers in this special section and also to the reviewers who helped us evaluate the manuscripts. Thanks are also extended to Helen Meng, Li Deng, and Kathy Jackson for their advice and assistance throughout the process.

DONG YU, *Lead Guest Editor*
Microsoft Research
Redmond, WA 98052 USA

GEOFFREY HINTON, *Guest Editor*
University of Toronto
Toronto, ON M5S 3G4, Canada

NELSON MORGAN, *Guest Editor*
International Computer Science Institute
Berkeley, CA 94704 USA

JEN-TZUNG CHIEN, *Guest Editor*
National Cheng Kung University
Tainan 701, Taiwan

SHIGEKI SAGAYAMA, *Guest Editor*
University of Tokyo
Tokyo 153-8902, Japan

Digital Object Identifier 10.1109/TASL.2011.2173371

Dong Yu (M'97–SM'06) received the Ph.D. degree in computer science from the University of Idaho, Moscow.

He is a Researcher in the Microsoft Speech Research Group, Microsoft Research, Redmond, WA. His current research interests include speech processing, robust speech recognition, discriminative model and training, voice search technology, and machine learning. He has over 80 publications in these areas, and is the inventor/co-inventor of more than 40 awarded and pending patents.

Dr. Yu is currently serving as an Associate Editor of the *IEEE Signal Processing Magazine*.

Geoffrey Hinton received the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 1978.

He spent five years as a faculty member at Carnegie-Mellon University, Pittsburgh, PA, and he is currently the Raymond Reiter Distinguished Professor of Artificial Intelligence at the University of Toronto, Toronto, ON, Canada, where he directs the program on Neural Computation and Adaptive Perception funded by the Canadian Institute for Advanced Research.

Prof. Hinton is a fellow of the Royal Society and an honorary foreign member of the American Academy of Arts and Sciences. His awards include the David E. Rumelhart Prize, the International Joint Conference on Artificial Intelligence research excellence award, and the Gerhard Herzberg Canada Gold Medal for Science and Engineering. He was one of the researchers who introduced the back-propagation algorithm. His other contributions include Boltzmann machines, distributed representations, time-delay neural nets, mixtures of experts, variational learning, contrastive divergence learning, and deep belief nets.

Nelson Morgan (S'76–M'80–SM'87–F'99) received the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1980.

He is the Director of the International Computer Science Institute, Berkeley, CA, where he has worked on speech processing since 1988. He is a Professor-in-Residence in the Electrical Engineering Department, University of California, Berkeley. In previous incarnations, he worked on EEG signal processing at the EEG Systems Laboratory and on speech analysis and synthesis at National Semiconductor. He has over 200 publications including three books, the most recent of which is a text on speech and audio processing coauthored with signal processing pioneer Ben Gold (a new revision of which is now being prepared with coauthor Dan Ellis of Columbia). His current interests include the incorporation of insights from studies of auditory cortex to practical speech recognition algorithms.

Prof. Morgan is a former and returning member of the IEEE Signal Processing Society Spoken Language Technical Committee, a former editor-in-chief of *Speech Communication* (and currently on its editorial board), and is now on the editorial board of the *IEEE Signal Processing Magazine*. He is a member of the International Speech Communication Association and is on its Advisory Council. In 1997, he received the Signal Processing Magazine best paper award (jointly with H. Bourlard).

Jen-Tzung Chien (S'97–A'98–M'99–SM'04) received the Ph.D. degree in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 1997.

Since 1997, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, where he is currently a Professor. He held the Visiting Researcher positions at Panasonic Technologies, Inc., Santa Barbara, CA, Tokyo Institute of Technology, Tokyo, Japan, Georgia Institute of Technology, Atlanta, GA, Microsoft Research Asia, Beijing, China, and IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include machine learning, speech recognition, face recognition, information retrieval, and blind source separation.

Dr. Chien is a senior member of the IEEE Signal Processing Society. He serves on the editorial board of the IEEE Signal Processing Letters. He received the Young Investigator Award (Ta-You Wu Memorial Award) from the National Science Council (NSC), Taiwan, in 2003, the Research Award for Junior Research Investigators from Academia Sinica, Taiwan, in 2004, and the NSC Distinguished Research Awards, in 2006 and 2010.

Shigeki Sagayama (M'82) received B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998 respectively, all in mathematical engineering and information physics.

He joined the Electrical Communications Laboratories, Nippon Telegraph and Telephone Public Corporation, in 1974. From 1990 to 1993, he was responsible for the Speech Processing Department at ATR Interpreting Telephony Research Laboratories and, from 1993 to 1998, he was with Speech and Acoustics Laboratory, NTT Human Interface Laboratories. In 1998, he became a Professor at the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. Since 2000, he has been a Professor at the Graduate School Information Science and Technology, University of Tokyo, Tokyo, Japan. His research interests include music signal processing, acoustic signal processing, music information processing, speech recognition and synthesis, signal processing, natural language processing, pattern recognition, and human interfaces. He has authored or coauthored in more than 480 technical publications and more than 80 patent applications.

Prof. Sagayama received the National Invention Award from the Institute of Invention, Japan, in 1991, the Technology Development Award from the Acoustical Society of Japan, in 1994, the Yamashita Memorial Research Award from the Information Processing Society, Japan, in 1995, the Paper Award from the Institute of Electronics, Information, Communications Engineers, Japan, and the Chief Official Award for Research Achievement from the Science and Technology Agency, Japan, both in 1996, and the Paper Award from the Information Processing Society of Japan in 2005. He is a member of the Acoustical Society of Japan, the Institute of Electronics, Information, and Communications Engineers, Japan, the Information Processing Society, Japan, and ESCA (European Speech Communication Association).