Introduction to the Issue on Statistical Learning Methods for Speech and Language Processing

IN the past few years, significant progress has been made in both research and commercial applications of speech and language processing. However, there remain important theoretical issues to be addressed on statistical modeling and learning. Theoretical advancement is expected to drive greater system performance improvement, which in turn generates the new need of novel learning and modeling methodologies as well as in-depth studies of them.

The main goal of the special issue that we have assembled here is to fill in the above need, with the focus on the fundamental modeling and learning issues of new emerging approaches and empirical applications in speech and language processing. Another focus of this special issue is on the cross-fertilization of learning approaches to speech and language processing problems. Many problems in speech and language processing share similarities (despite some conspicuous differences), and techniques in these two fields can be successfully cross-pollinated. Our additional goal is to bring together a diverse but complementary set of contributions on emerging learning methods for speech processing, language processing, as well as unifying approaches to problems cross cutting these two fields.

Discriminative learning has become a major theme in most areas of speech and language processing. One of the recent advances in discriminative learning is the integration of the large margin idea, which is the classical training standard in machine learning, into the conventional discriminative training criteria for string recognition. In the first paper, Heigold *et al.* discuss how typical training criteria, such as minimum phone error and maximum mutual information, can be extended to incorporate the margin concept. In this work, a new margin-based formalism is proposed for various conventional training criteria. Experimental results show that the new criteria help the performance across a wide variety of string recognition scenarios including speech recognition, concept tagging, and handwriting recognition. In another paper, Cheng *et al.* explore online learning and acoustic feature adaptation in large margin hidden Markov models (HMMs), which lead to a better optimization method for large-margin HMM training. Moving beyond acoustics, language modeling is one of the essential problems in speech and language fields. Zhou *et al.* introduce a novel pseudo-conventional N-gram language model with discriminative training, and also carry out an empirical study of the robustness of discriminatively trained LMs. Experimental results show that cumulative performance improvements can be achieved via this method.

Sequential pattern classification is at the core of many speech and language processing problems. Conditional random field (CRF) is a widely adopted approach to supervised sequential labeling. However, the computational load and model complexity grow dramatically when taking complex structure into account. Here, Sokolovska *et al.* address this issue through efficient feature selection based on imposing sparsity through an L1 regularization for CRF. The results show that, without performance degradation, the L1 regularized CRF results in significantly faster training and labeling speed, and hence makes it possible to scale up systems to handle very large dimensional models. Meanwhile, Yu *et al.* improve the CRF model from another perspective. They proposed a multi-layer sequence classification algorithm where each layer is a CRF, and each higher layer's input consists of both the previous layer's observation sequence and the resulting frame-level marginal probabilities. Compared with the conventional CRF, the deep-structured CRF achieves superior labeling accuracy on common tagging tasks. Using the kernel method to improve the performance of sequential pattern classifiers is also an important direction. Kubo *et al.* describe a novel sequential pattern classifier based on kernel methods. Unlike conventional approaches, they use kernel methods to estimate the emission probability of HMM, with the extra benefit due to the powerful nonlinear classification capability of kernel methods. On the other hand, unlike conventional CRF/HMM-based methods, Bellegarda attacks this problem from a novel angle based on latent semantic mapping and obtains insightful results.

In many speech and language applications, machine learning technologies play a critical role. This issue collects some latest advances of machine learning techniques in speech and language processing. The HMM is a widely adopted model. However, standard HMMs have severe limitations when they are applied to speech recognition. To accommodate these problems, Nguyen and Zweig propose flat direct models, where the posterior distribution of a sequence of words is directly modeled with no inherent notion of word order or local contiguous statistical dependency. Ensemble learning is also an important topic in machine learning, and has many successful applications in the speech and language processing area. In their paper, Shinozaki *et al.* propose unsupervised cross-validation and aggregated adaptation algorithms that integrate the ideas of ensemble methods to adapt acoustic models and give superior results. Wolfe *et al.* propose a likelihood-based semi-supervised approach for model selection and discuss its speech-related applications, especially pronunciation selection for un-transcribed spoken words. Superior results are reported on a speech recognition task. The paper by Mitra *et al.* deals with a rather specific area of speech processing, speech inversion. In their work, a set of machine learning strategies is compared in the scenario of vocal tract variable retrieval.

Exploring unified modeling approaches across the speech and language processing area is a main theme of this special issue. As an example, HMM has been successfully applied to both speech recognition and speech synthesis. However, Dines *et al.*

Digital Object Identifier 10.1109/JSTSP.2010.2086910

show that, despite essentially the same statistical model, the optimal systems for automatic speech recognition (ASR) and text-to-speech (TTS) are often very different in many aspects. This demonstrates one of the many challenges in the investigation of unified modeling approaches.

This issue also collects latest advances on several important speech and language processing tasks, e.g., speaker diarization of conversations is an interesting problem and task in speech processing. Kenny *et al.* present comprehensive comparison of three systems for the speaker diarization problem, and proposed a method of integrating the eigen-voice and eigen-channel priors with the variational Bayes-based diarization approach. Voice activity detection is an essential problem in many real world speech systems such as speech recognition. Cournapeau *et al.* present an online method for this problem. The authors propose a variational Bayesian framework-based method which leads to significant performance improvement. Finally, language recognition is an important task in speech and language processing. Gonzalez *et al.* discuss their state-of-the-art ATVS-UAM system for the NIST 2009 Language Recognition Evaluation. Three systems and a combination method are presented. Sufficient details of the systems are provided for readers who are interested.

In summary, we hope that this special issue has achieved its goal of underlining the importance of statistical modeling and learning for speech and language processing, with its special focus on cross fertilization among speech processing, language processing, and machine learning. Enjoy reading the papers.

We would like to thank the authors for their quality submissions and the reviewers for their time and effort in the review process. We are grateful to the former and the current Editor-in-Chief, Prof. Lee Swindlehurst and Prof. Vikram Krishnamurthy, for their encouragement and support. Finally, we also want to thank Rebecca Wollman and Jayne Huber for their assistance in assembling this special issue.

XIAODONG HE, *Lead Guest Editor*

Microsoft Research
Redmond, WA 98052 USA
(e-mail: xiaohe@microsoft.com)

LI DENG, *Guest Editor*

Microsoft Research
Redmond, WA 98052 USA
(e-mail: deng@microsoft.com)

ROLAND KUHN, *Guest Editor*

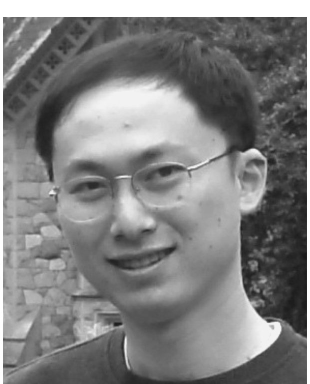
National Research Council of Canada
Gatineau, QC J8X 3X7 Canada
(e-mail: roland.kuhn@cnrc-nrc.gc.ca)

HELEN MENG, *Guest Editor*

The Chinese University of Hong Kong
Hong Kong, China
(e-mail: hmmeng@se.cuhk.edu.hk)

SAMY BENGIO, *Guest Editor*

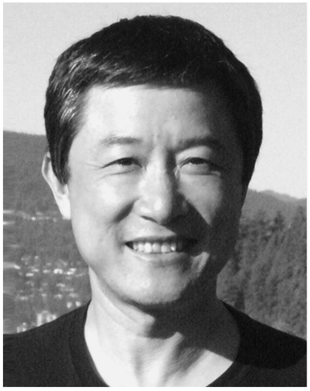
Google, Inc.
Mountain View, CA 94043 USA
(e-mail: bengio@google.com)



Xiaodong He (M'03–SM'08) received the B.S. degree from Tsinghua University, Beijing, China, in 1996, the M.S. degree from the Chinese Academy of Sciences, Beijing, in 1999, and the Ph.D. degree from the University of Missouri-Columbia in 2003.

He joined the Speech and Natural Language Group, Microsoft Corporation, in 2003, and the Natural Language Processing Group, Microsoft Research, Redmond, WA, in 2006, where he currently serves as a Researcher. His research areas include statistical machine learning, automatic speech recognition, natural language processing, machine translation, and human–computer interaction. In these areas, he has authored/coauthored more than 30 refereed papers in leading international conferences and journals and has filed more than ten patents. He coauthored the book *Discriminative Learning for Speech Recognition: Theory and Practice* (Morgan and Claypool, 2008). He and colleagues developed the machine translation system that achieved the best result in the Chinese-to-English constrained training track of the 2008 NIST MT Open Evaluation.

Dr. He served as co-chair of the NIPS 2008 workshop on Speech and Language: Learning Based Methods and Systems. He also served on program committees of various conferences in the areas of speech recognition, natural language processing, machine learning, and pattern recognition. He is a member of ACL and a member of Sigma Xi.



Li Deng (M'86–SM'91–F'05) received the B.S. degree from the University of Science and Technology of China, Hefei, (with the Guo Mo-Ruo Award), and the Ph.D. degree from the University of Wisconsin-Madison (with the Jerzy E. Rose Award).

In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada as an Assistant Professor, where he became a Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as a Senior Researcher, where he is currently a Principal Researcher. He is also an Affiliate Professor in the Department of Electrical Engineering, University of Washington, Seattle. His past and current research activities include automatic speech and speaker recognition, spoken language identification and understanding, speech-to-speech translation, machine translation, statistical methods and machine learning, neural information processing, deep-structured learning, machine intelligence, audio and acoustic signal processing, statistical signal processing and digital communication, human speech production and perception, acoustic phonetics, auditory speech processing, auditory physiology and modeling, noise robust speech processing, speech synthesis and enhancement, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 300 refereed papers in leading international conferences and journals, 12 book chapters, and has given keynotes, tutorials, and lectures worldwide. He has been granted over 30 U.S. or international patents in acoustics, speech/language technology, and signal processing.

Dr. Deng is a Fellow of the Acoustical Society of America. He authored or coauthored three books in speech processing and learning. He serves on the Board of Governors of the IEEE Signal Processing Society (2008–2010), and as Editor-in-Chief for the IEEE SIGNAL PROCESSING MAGAZINE (2009–2012), which ranks consistently among the top journals with the highest citation impact. According to the Thomson Reuters Journal Citation Report, released June 2010, the SPM has ranked first among all IEEE publications (125 in total) and among all publications within the Electrical and Electronics Engineering Category (245 in total) in terms of its impact factor.

Roland Kuhn (M'92) received the Ph.D. degree in computer science from McGill University, Montreal, QC, Canada, in 1993

He is a Research Officer with the National Research Council of Canada (NRC), Gatineau. After the Ph.D. degree, he worked for the Centre de Recherche Informatique de Montréal (CRIM) until September 1996. Subsequently, he worked for the Panasonic Speech Technology Laboratory, Santa Barbara, CA (October 1996–June 2004). During this first period of his career, his research focused on areas related to speech: automatic speech recognition, speaker adaptation, dialogue, and speaker verification/identification. In July 2004, he joined NRC and embarked on research in machine translation. He has authored 44 refereed publications and holds 29 U.S. patents; his best-known research contributions are the cache language model and eigenvoices for speaker adaptation.

Dr. Kuhn was a member of the IEEE Speech Technical Committee from 2002 to 2004.

Helen Meng (M'98–SM'09) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge.

She joined The Chinese University of Hong Kong in 1998, where she is currently Professor in the Department of Systems Engineering and Engineering Management. In 1999, she established the Human–Computer Communications Laboratory and serves as Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies and serves as Co-Director. This laboratory was recognized as a Ministry of Education of China (MoE) Key Laboratory in 2008. She also served as an Associate Dean (Research) of the Faculty of Engineering from 2006 to 2010.

Prof. Meng received the Higher Education Outstanding Scientific Research Output Awards in Technological Advancements, for the area of "Multimodal User Interactions with Multilingual Speech and Language Technologies." in 2009 Her research interest is in the area of human–computer interaction via multimodal and multilingual spoken language systems, computer-aided language learning systems, as well as translingual speech retrieval technologies. She serves as Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. She is also an elected board member of the International Speech Communication Association.

Samy Bengio (M'04) received the Ph.D. degree in computer science from the University of Montreal, Montreal, QC, Canada, in 1993.

He has been a Research Scientist at Google, Inc., Santa Barbara, CA, since 2007. Before that, he was a Senior Researcher in statistical machine learning at IDIAP Research Institute, Martigny, Switzerland, where he supervised Ph.D. students and postdoctoral fellows. His research interests span many areas of machine learning such as support vector machines, time series prediction, mixture models, large-scale problems, speech recognition, multi-channel and asynchronous sequence processing, multi-modal (face and voice) person authentication, brain–computer interfaces, and document retrieval. He is an Associate Editor of the *Journal of Computational Statistics* and on the editorial boards of the *Journal of Machine Learning Research* and the *Machine Learning Journal*.

Dr. Bengio has been general chair of the Workshops on Machine Learning for Multimodal Interactions (MLMI'2004, 2005, and 2006), Program Chair of the IEEE Workshop on Neural Networks for Signal Processing (NNSP'2002), and on the program committee of several international conferences such as NIPS, ICML, ECML, and IJCAI. More information can be found on his website: http://bengio/abracadoudou.com.