Editorial
Special Section on Statistical and Perceptual Audio Processing

HUMAN perception has always been an inspiration for automatic processing systems, not least because tasks such as speech recognition only exist because people do them—and, indeed, without that example we might wonder if they were possible at all. As computational power grows, we have increasing opportunities to model and duplicate perceptual abilities with greater fidelity, and, most importantly, based on larger and larger amounts of raw data describing both what signals exist in the real world, and how people respond to them.

The power to deal with large data sets has meant that approaches that were once mere theoretical possibilities, such as exhaustive search of exponentially-sized codebooks, or real-time direct convolution of long sequences, have become increasingly practical and even unremarkable. A major consequence of this is the growth of statistical or corpus-based approaches, where complex relations, discriminations, or structures are inferred directly from example data (for instance by optimizing the parameters of a very general algorithm). An increasing number of complex tasks can be given empirically optimal solutions based on large, representative datasets.

The traditional idea of perceptually-inspired processing is to develop a machine algorithm for a complex task such as melody recognition or source separation through inspiration and introspection about how individuals perform the task, and on the basis of direct psychological or neurophysiological data. The results can appear to be at odds with the statistical perspective, since perceptually-motivated work is often *ad-hoc*, comprising many stages whose individual contributions are difficult to separate.

We believe that it is important to unify these two approaches: to employ rigorous, exhaustive techniques taking advantage of the statistics of large data sets to develop and solve perceptually-based and subjectively-defined problems. With this in mind, we organized a one-day workshop on Statistical and Perceptual Audio Processing as a satellite to the International Conference on Spoken Language Processing (ICSLP-INTERSPEECH), held in Jeju, Korea, in September 2004. A second workshop will be held at the next ICSLP-INTERSPEECH in Pittsburgh, PA, in September 2006.

Although independent of the workshops, this special issue is based on the same insight and goal. In our call for papers, and in our editorial choices of which papers to consider for the special issue, we were looking specifically for rigorous, statistical techniques applied to perceptually-defined tasks or processing in innovative and interesting ways. As can be seen from the papers that follow, this seemingly narrow focus can be interpreted in a wide variety of ways. In fact, we received more than 50 submissions for the special issue—many more than we expected—which says something about the timeliness and resonance of this idea.

About half of the papers actually fit our idea of a genuine combination of statistical and perceptual techniques (the remainder were forwarded as regular submissions to the TRANSACTIONS). The ten papers eventually accepted in time for this issue provide a satisfyingly broad range of interpretations of both "statistical" and "perceptual."

In terms of the perceptual aspects, some papers addressed tasks that were strongly based in perception and cognition, for instance automatically predicting the "mood" of music, as considered by Lu, Liu, and Zhang. At the other extreme, several authors took rather precise perceptual models of signal masking, such as the ones used in MPEG audio encoders, and applied them to tasks such as speech enhancement (Ma, Bouchard, and Goubran) and echo cancellation in telephony (Gordy and Goubran).

The neurophysiological basis of perception was the basis of the model of Holmberg, Gelbart, and Hemmert, who investigate the benefits of modeling peripheral neural activity in a speech recognizer. Blumensath and Davies also employ a neural motivation, at least at an abstract level, for their sparse coding schemes for audio modeling. The very specific mechanisms by which humans and other animals can perceive spatial position so well using just two ears motivated the binaural mask-based speech recognition system of Harding, Barker, and Brown.

Clearly, there was also a broad range of topics addressing the statistical side. Tasks include recognition, e.g., of instruments in mixtures (Essid, Richard, and David), and estimating perceived similarity of musical timbre (Mörchen, Ultsch, Thies, and Löhken). Considering more classical tasks in signal processing from a perceptual perspective are Vincent, looking at source separation, and Christensen and Jensen, who apply perceptual criteria to sinusoidal signal models.

Overall, we are very pleased that the idea of uniting and founding perceptual-style processing in a firm statistical basis has proven to be appealing to authors, and so broad in its applicability. We look forward to an increasing body of high-quality research aligned to this idea. Finally, we are very much indebted to Isabel Trancoso for giving us the opportunity to edit this special issue, to Kathy Jackson at the IEEE for patiently working through our collective ignorance of the appropriate procedures, and to more than 70 reviewers who diligently and thoughtfully helped to shape the papers, all on a tight time line. We believe the results justify the effort, and hope you will agree.

Digital Object Identifier 10.1109/TSA.2005.862700

DANIEL P. W. ELLIS, *Guest Editor*
Columbia University
Electrical Engineering Department
New York, NY 10027 USA
dpwe@ee.columbia.edu

BHIKSHA RAJ, *Guest Editor*
Mitsubishi Electric Research Labs
Cambridge, MA 02138 USA
bhiksha@merl.com

JUDITH C. BROWN, *Guest Editor*
Wellesley College
Physics Department
Wellesley, MA 02101 USA
jbrown@wellesley.edu

MALCOLM SLANEY, *Guest Editor*
Yahoo! Research Laboratory
Sunnyvale, CA 94088 USA
malcolm@ieee.org

PARIS SMARAGDIS, *Guest Editor*
Mitsubishi Electric Research Labs
Cambridge, MA 02138 USA
paris@merl.com

Daniel P. W. Ellis (S'92–M'96–SM'04) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, where he was a Research Assistant in the Media Lab.

He spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA. Currently, he is an Associate Professor with the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY email list of 1700 worldwide researchers in perception and cognition of sound.

Bhiksha Raj received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in May 2000.

Since 2001, he has been with Mitsubishi Electric Research Laboratories, Cambridge, MA. He is working mainly on algorithmic aspects of speech recognition, with special emphasis on improving the robustness of speech recognition systems to environmental noise. His professional interests include robust speech recognition, computational auditory scene analysis, statistical audio processing, microphone array processing, and data mining.

Judith C. Brown is Professor Emeritus of physics at Wellesley College, Wellesley, MA, and Visiting Scientist at the Media Lab of the Massachusetts Institute of Technology, Cambridge. Her interests lie in signal processing of musical sounds, and more recently classification of vocalizations of marine mammals.

Prof. Brown is a Fellow of the Acoustical Society of America and has served on its Technical Committee on Musical Acoustics for over ten years.

Malcolm Slaney (S'84–M'01–SM'01) received his B.S., M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN.

He is a Senior Research Scientist at Yahoo! Research Laboratory, where he is spearheading efforts in multimedia signal processing research. He has been employed at Bell Laboratories, Schlumberger Palo Alto Research, Apple's Advanced Technology Group, Interval Research, and IBM's Almaden Research Center. His work spans the fields of multimedia (indexing and segmentation), machine learning (combining information from multiple sources), morphing (audio and video), computer graphics (facial synthesis from video databases), tomography, perception (pitch, timbre, and computational auditory scene analsys) and user interfaces. He is an author (with A. C. Kak) of the book *Principles of Computerized Tomographic Imaging*, which was recently republished by SIAM as one of their Classics in Applied Mathematics. He is an editor (with Steven Greenberg) of *Computational Models of Auditory Function*. He is an instructor at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA, and has organized the Hearing Seminar for the last 15 years.

Paris Smaragdis (M'03) received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, where he completed his graduate and postdoctoral training.

He is a Member of the Research Staff at Mitsubishi Electric Research Laboratories, Cambridge. His interests are computational audition, scene analysis and the intersection of machine learning with signal processing.