Introduction to the Issue on Speech Processing for Natural Interaction With Intelligent Environments

Introduction to the Issue on Speech Processing for Natural Interaction With Intelligent Environments

WITH the advances in microelectronics, communication technologies, and smart materials, our environments are transformed to be increasingly intelligent by the presence of robots, bio-implants, mobile devices, advanced in-car systems, smart house appliances, and other professional systems. As these environments are integral parts of our daily work and life, there is a great interest in a natural interaction with them. Also, such interaction may further enhance the perception of intelligence. "Interaction between man and machine should be based on the very same concepts as that between humans, i.e., it should be intuitive, multi-modal and based on emotion.", as envisioned by Reeves and Nass (1996) in their famous book *The Media Equation*. Speech is the most natural means of interaction for human beings and it offers the unique advantage that it does not require carrying a device for using it since we have our "device" with us all the time.

Speech processing techniques are therefore developed to support either explicit interaction through message communications, or implicit interaction by providing valuable information about the physical ("who speaks when and where") as well as the emotional and social context of an interaction. However, intelligent environments are characterized with distant microphone(s), resource constraints, and large variations in acoustic condition, speaker, content, and context, all being significant challenges to speech processing.

To achieve the goal of natural interaction, a broad range of topics are to be addressed. We roughly group them into four clusters: 1) multi-microphone front-end processing and joint optimization with automatic speech recognition (ASR), 2) ASR in adverse acoustic environments and for low-resource and distributed computing infrastructure, 3) speaker diarization, affective computing and context awareness for interaction, and 4) cross-modal analysis of audio and visual data for smart spaces.

Much progress has been made, yet there is still a long way to go. More importantly, synergy across fields is demanded. This issue aims to bring together researchers and engineers to present latest developments in the given fields in one place and hopefully will stimulate cross-fertilization. The clusters of topics are well represented by the 12 papers in this issue and certainly, several papers fall into multiple clusters.

The first two papers deal with microphone array speech processing. Yoon *et al.* incorporate an acoustic model combination method with a hidden Markov model (HMM)-based mask estimation method for multichannel source separation. The acoustic model combination method is used to reduce the mismatch between training and testing conditions of an ASR after applying source separation. The paper by Yu and Hansen presents a time–frequency domain blind beamforming approach, without a prior knowledge of the array shape, for extracting the desired speech from a noisy musical environment, along with the evaluations conducted in both real in-vehicle and simulated noisy environments.

The succeeding five papers are concerned with the robustness to environmental distortions and reverberation in single-channel scenarios, and the constraints resulting from limited computational resources available in embedded devices. The paper by Tan and Lindberg addresses both issues by proposing a frame selection algorithm which emphasizes the reliable regions, thus improving the ASR robustness. As hardly any frame is selected for non-speech regions, the method may also serve as a robust voice activity detector (VAD) and a scalable source coding scheme for distributed speech recognition (DSR). Borgström and Alwan present a low complexity algorithm for determining improved speech presence probabilities using HMM-based inference to exploit the temporal correlation present in spectral speech data. The algorithm is applied to soft-decision enhancement and further to noise-robust ASR. The contribution by Ichikawa *et al.* covers robustness towards reverberation resulting from the large distance between the source and the microphone. The authors propose to compute dynamic features in the linear-logarithmic hybrid domain for distant-talking ASR in reverberant environments. The paper by Astudillo *et al.* presents an uncertainty propagation approach for the advanced front-end of the ETSI DSR standards. The advantage of the method lies in the fact that the uncertainty is determined in the domain where most speech enhancement methods operate by using self similarity measures. Fukuda *et al.* present a statistical-model based noise-robust VAD algorithm using long-term temporal information and harmonic-structure based features in speech. The algorithm works well both as a VAD and a robust preprocessing method for ASR.

The three papers in the third cluster address the problem of context acquisition from audio or audio-visual signals. Schmalenstroeer and Haeb-Umbach present a low-latency diarization system for smart homes that conducts joint speaker segmentation, localization, and identification, supported by face identification. Speaker positions obtained though a blind beamforming method are combined with speaker change information to improve speaker identification. Stafylakis *et al.* derive a new approach to the Bayesian information criterion (BIC), which combines the strengths of the global and local BIC, and apply the resulting segmental BIC for speaker diarization. The contribution by Wöllmer *et al.* incorporates long short-term memory and dynamic Bayesian networks for incremental recognition of the user's emotional space. The automatic estimation of human affect from the speech signal makes virtual agents more natural and human-like.

The final two papers explore cross-modal analysis. Shivappa *et al.* develop a scheme to fuse audio and visual cues to track multiple persons in an intelligent meeting room equipped with

Digital Object Identifier 10.1109/JSTSP.2010.2069910

multiple cameras and microphone arrays. The scheme performs comparably to particle filters and is robust to calibration errors. Tracking is a key step towards natural interaction. The paper by Naqvi *et al.* presents a multimodal approach to blind source separation (BSS) of moving sources. A full 3-D visual tracker based on particle filtering is implemented to provide velocity and direction of sources; a beamforming algorithm is used when sources are moving and a BSS algorithm is performed when sources are stationary.

We would like to thank the authors for submitting quality papers and our reviewers for their thoughtful and timely reviews. We also thank the former Editor-in-Chief of this journal, Prof. Lee Swindlehurst, for his encouragement and support. Finally, we are grateful to Jayne Huber and Rebecca Wollman for their assistance in assembling this issue.

ZHENG-HUA TAN, *Lead Guest Editor*
Department of Electronic Systems
Aalborg University
Aalborg 9220, Denmark
zt@es.aau.dk

REINHOLD HAEB-UMBACH, *Guest Editor*
Department of Communications Engineering
University of Paderborn
D-33098 Paderborn, Germany
haeb@nt.uni-paderborn.de

SADAOKI FURUI, *Guest Editor*
Department of Computer Science
Tokyo Institute of Technology
Tokyo 152-8552, Japan
furui@cs.titech.ac.jp

JAMES R. GLASS, *Guest Editor*
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
glass@mit.edu

MAURIZIO OMOLOGO, *Guest Editor*
SHINE Research Unit
FBK-IRST
38050 Povo-Trento, Italy
omologo@fbk.eu

Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999.

He is an Associate Professor in the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, which he joined in May 2001. Prior to that, he was a Postdoctoral Fellow in the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea, and an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University. His research interests include speech recognition, noise robust speech processing, multimedia signal and information processing, multimodal human–computer interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He edited the book *Automatic Speech Recognition on Mobile Devices and over Communication Networks* (Springer, 2008). He serves as an Editorial Board Member for *Elsevier Computer Speech and Language*, and the *International Journal of Data Mining, Modeling, and Management*.

Reinhold Haeb-Umbach (M'89) received the Dipl.-Ing. and Dr.-Ing. degree in electrical engineering from RWTH Aachen University, Aachen, Germany, in 1983 and 1988, respectively.

From 1988 to 1989, he was a Postdoctoral Fellow at the IBM Almaden Research Center, San Jose, CA, conducting research on coding and signal processing for recording channels. From 1990 to 2001, he was with Philips Research working on various aspects of automatic speech recognition, such as acoustic modeling, efficient search strategies, and mapping of algorithms on low-resource hardware. Since 2001, he has been a Professor in communications engineering at the University of Paderborn, Paderborn, Germany. His main research interests are in statistical speech signal processing and recognition and in signal processing for communications. He has published more than 100 papers in peer reviewed journals and conferences.

Sadaoki Furui (M'79–SM'88–F'93) received the B.S., M.S., and Ph.D. degrees from Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Nippon Telegraph and Telephone Corporation (NTT) Labs in 1970, he has worked on speech analysis, speech recognition, speaker recognition, speech synthesis, speech perception, and multimodal human–computer interaction. From 1978 to 1979, he was a Visiting Researcher at AT&T Bell Laboratories, Murray Hill, NJ. He was a Research Fellow and the Director of the Furui Research Laboratory, NTT Labs, and is currently a Professor at the Department of Computer Science, Tokyo Institute of Technology. He has authored or coauthored over 800 published papers and books including *Digital Speech Processing, Synthesis, and Recognition* (Marcel Dekker, 1989).

Prof. Furui received the Paper Award and the Achievement Award from the Institute of Electronics, Information, and Communication Engineers of Japan (IEICE) (1975, 1988, 1993, 2003, 2003, 2008), and the Paper Award from the Acoustical Society of Japan (ASJ) (1985, 1987). He received the Senior Award and Society Award from the IEEE Signal Processing Society (1989, 2006), the International Speech Communication Association (ISCA) Medal for Scientific Achievement (2009), and the IEEE James L. Flanagan Speech and Audio Processing Award (2010). He also received the Achievement Award from the Minister of Science and Technology and the Minister of Education, Japan (1989, 2006), and the Purple Ribbon Medal from Japanese Emperor (2006).

James Glass (M'78–SM'06) received the B.Eng. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 1982, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology MIT, Cambridge, in 1985, and 1988, respectively.

He is a Principal Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) where he heads the Spoken Language Systems Group. He is also a Lecturer in the Harvard–MIT Division of Health Sciences and Technology. After starting in the Speech Communication Group at the MIT Research Laboratory of Electronics, he has worked since 1989 at the Laboratory for Computer Science, and since 2003 at CSAIL. His primary research interests are in the area of speech communication and human–computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, supervised students, and published extensively in these areas.

Dr. Glass is currently a member of the IEEE Signal Processing Society Speech and Language Processing Technical Committee, an ISCA Distinguished Lecturer, an associate editor for *Computer, Speech, and Language*, and the *EURASIP Journal on Audio, Speech, and Music Processing*, and has been a past Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Maurizio Omologo (M'88) was born in Padua, Italy, in 1959. He received the "Laurea" degree (with honors) in electrical engineering from the University of Padua, Padua, Italy, in 1984.

From 1984 to 1987, he was a Researcher in speech coding at CSELT, Torino, Italy. In 1988, he joined ITC-IRST (now Fondazione Bruno Kessler-IRST) Trent, Italy, where he is the head of the SHINE (Speech-acoustic Scene Analysis and Interpretation) research unit. He has also been teaching audio signal processing at the University of Trento since 2001. His current research interests include audio and speech processing, acoustic scene analysis, and automatic speech recognition, in particular for distant-talking scenarios. He is the author of more than 100 papers in major international conferences and journals in the field.

Dr. Omologo served as an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2003 to 2005. He is currently an editorial board member of the *Language Resources and Evaluation* journal. He served as General Co-Chairman of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2001, of Hands-free Speech Communication and Microphone Arrays (HSCMA) 2008, and Local Chair of IEEE-ASRU 2009. Between 2006 and 2009, he acted as Project Manager of the DICIT (Distant-talking Interfaces for Control of Interactive TV) European Project.