Introduction to the Special Issue on Multimodal Processing in Speech-Based Interactions

SPEECH constitutes the primary form of human communication. However, traditional automatic speech processing based on information from the audio channel alone deviates from the way humans interact by ignoring information available in additional modalities, for example the visual channel. Therefore, not surprisingly, audio-only automatic speech processing and interaction lag in performance, robustness, and naturalness to what humans achieve. Acknowledging these shortcomings, there has been increasing research interest in mimicking the human capacity to *jointly* process information available in multiple modalities related to human activities. As part of the resulting efforts, technological developments in individual modalities for human–computer interaction are being tirelessly extended to incorporate multimodal processing. For example, “single mode” speech recognition is migrating towards audiovisual speech recognition that exploits information from the speaker’s facial and lip motions; traditional text-to-speech synthesis is being extended to also synthesize video of the speaker’s head, facial, and lip motions; speaker recognition is becoming multimodal by including additional biometric traits such as facial static images or videos and fingerprints; unimodal speech-based interfaces are migrating to multimodal ones by including pen and gestural input; and databases for speech technology evaluation are migrating from audio-only ones (for example telephony speech) towards multimodal, multisensory recordings (for example of interaction in meeting rooms), enabling the development and evaluation of multimodal speech technologies. In parallel, we are also witnessing the emergence of major programs with focus on the subject, such as the “Rich Transcription evaluation” of recorded meetings overseen by the National Institute of Standards and Technology in the U.S., as well as the strategic objective of the Information Society Technologies Program of the European Union to develop “multimodal interfaces.”

The above trends have motivated us to proceed with this special issue. Multimodality in speech is, of course, a rather broad topic that poses significant challenges to overcome. One is due to its very multidisciplinary nature—clearly, processing diverse modalities and extracting reliable information from each of them requires algorithms from different fields; for example, computer vision and signal processing. Another issue has to do with limited appropriate database resources, compared to what is available for the traditional acoustic speech modality. Above all, a critical factor to the effectiveness of multimodal speech processing is the robust integration, or fusion, of information from the various modalities. These topics are successfully covered in this special issue that contains nine papers. The majority of the contributions (six papers) focus on audiovisual speech, in particular the problems of speech recognition, synthesis, and inversion. Two additional papers discuss multimodal interaction using speech and pen input, and a ninth one addresses modeling group interaction based on audiovisual input during meetings.

A central theme that occupies much of the special issue is that of multimodal fusion, with application to audiovisual speech inversion, recognition, separation, and synthesis. In particular, Katsamanis *et al.* propose an adaptive piece-wise linear model that employs audiovisual speech observations to estimate articulatory speech parameters using canonical correlation analysis for dimensionality reduction and late decision fusion schemes. In an accompanying paper by the same group, Papandreou *et al.* consider the audiovisual automatic speech recognition problem within the general framework of multimodal fusion, where they introduce the concept of adaptive uncertainty compensation to account for the varying reliability of the feature streams of the modalities of interest. The problem of fusion is also addressed in the contribution of Sánchez-Soto *et al.*, who provide theoretical results on stream weight estimation and propose estimates for stream reliability and informativeness. Barker *et al.* deviate from the above in that they propose a speech fragment decoding framework emphasizing the importance of visual speech not only for recognition, but also for speech separation, namely in the presence of competing nonstationary noise. Two papers on visual speech synthesis by Melenchón *et al.* and Tao *et al.* address problems in audiovisual speech synthesis that have largely remained unsolved to date, namely modeling techniques for the effective synchronization between the audio and the visual channel, which is also useful to increase the level of expressiveness for visual speech synthesis systems in order to create emotions or emphasis that can be perceived through both audio and the video channels. Both papers also demonstrate the importance of suitable audiovisual databases for the development of practical systems.

Aside from the visual face modality, pen is also a key modality that can significantly enhance speech-based interactions. There are two papers in this special issue that explore the complementary relationship between the speech and pen modalities. Both papers apply dynamic programming to combine partial hypotheses across modalities. The paper by Liu *et al.* demonstrates efficient partial hypotheses fusion in an English dictation task, where the interface enables users to correct misrecognized words by handwriting. Another related application is Chinese character recognition, where syllable recognition generates a large number of homophonic characters and disambiguation is achieved by recognizing partial pen strokes of the intended character. The investigation by Hui *et al.* is based on a map navigation task with multiple referents in either or both modalities. Cross-modal hypotheses fusion enables robust interpretation of the referents in face of speech and pen recognition errors. Both papers show that combining speech and pen offers ease in entering complex user requests and commands. Intermodal relations such as complementarity

Digital Object Identifier 10.1109/TASL.2009.2013699

and redundancy can be leveraged in multimodal interpretation for improved accuracy, robustness, and efficiency.

Another topic of emerging importance is coprocessing of audio and visual information from multiple speakers during collaborative exchanges in order to extract meaningful social signaling information. The paper by Jayagopi *et al.* compares: 1) supervised versus unsupervised approaches to reliably classifying participants by their dominance level, and 2) the relative utility of tracking specific audio and visual information sources individually versus jointly. Promising work on this and related topics can offer important information for understanding social influence and leadership, diagnosing maladaptive social exchanges, and facilitating effective meetings and classroom interactions. One challenge that remains to be addressed in this area is the robust recognition of social signals in more naturalistic noisy environments. Considerably more user data will be required to determine which multimodal metrics can most reliably predict social dominance across a varied cross section of people and application contexts. Finally, further exploration will be required to identify the most valuable metrics for rapidly and reliably extracting social cues from a range of modalities, for example the visual information sources discussed in this paper.

It is our hope that the readers will find this special issue informative and interesting. We would like to first of all thank the authors of all submitted papers. Unfortunately, we could not include all contributed papers due to the aggressive schedule of the special issue. Further, we wish to offer our sincere thanks to our Editor-in-Chief, Professor Mari Ostendorf and the Publications Coordinator, Ms. Kathy Jackson, for their advice and assistance during the preparation of this issue.

HELEN MENG
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong, SAR, China

SHARON OVIATT
Incaa Designs,
Bainbridge Island, WA 98110 USA

GERASIMOS POTAMIANOS
Institute of Computer Science, Foundation for Research and Technology—Hellas (ICS-FORTH),
Heraklion, Crete GR-711 10, Greece

GERHARD RIGOLL
Institute for Human–Machine Communication, Munich University of Technology,
Munich 80333, Germany

Helen Meng received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology in 1995.

She is a Professor at Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, where she also currently serves as the Associate Dean of Research for the Faculty of Engineering. She started serving as Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING in January 2009. Her research interests include multilingual spoken dialog systems, spoken document retrieval, multimodal biometric authentication technologies, and multimodal user interfaces.

Sharon Oviatt is President and Chair of the Board of Directors of Incaa Designs, Bainbridge Island, WA (www.incaadesigns.org). She is internationally known for her work in human-centered interface design and evaluation, educational interfaces, mobile interfaces, and pen, speech, and multimodal interfaces. She has published over 130 scientific papers in a wide range of venues, many focusing on multimodal interfaces.

In 2000, she received a National Science Foundation Creativity Award for pioneering research on mobile multimodal interfaces.

Gerasimos Potamianos received the Ph.D. degree in electrical and computer engineering from The Johns Hopkins University (JHU), Baltimore, MD, in 1994.

Since then, he has worked at the Center for Language and Speech Processing, JHU, at AT&T Labs—Research, and at the IBM T. J. Watson Research Center. He currently holds a Research position at the Institute of Computer Science (ICS), FORTH, Heraklion, Greece. His research interests span the areas of multimodal speech processing with applications to human–computer interaction and ambient intelligence, with particular emphasis on audiovisual speech processing, automatic speech recognition, and multimedia signal processing and fusion.

Gerhard Rigoll received the Dr.-Ing. degree in automatic speech recognition from the Fraunhofer Institute, Stuttgart, Germany, in 1986 and the Dr.-Ing. habil. degree from Stuttgart University in 1991.

From 1986 to 1988, he was a Postdoctoral Fellow at the IBM T. J. Watson Research Center, Yorktown Heights, NY. From 1991 to 1993, he worked as a Guest Researcher for the NTT Human Interface Laboratories, Tokyo, Japan. In 1993, he was appointed a Full-Professor of computer science at Gerhard–Mercator–University, Duisburg, Germany. Since 2002, he has been heading the Institute for Human–Machine Communication, Munich University of Technology. His research interests are in the field of multimodal human–machine communication, covering areas such as speech and handwriting recognition, gesture and action recognition, face identification, emotion recognition, person tracking, and interactive computer graphics.