EURASIP Journal on Applied Signal Processing 2002:11, 1151–1153

Editorial

Chalapathy Neti
IBM T.J. Watson Research Center, Rte 134, Yorktown Heights, NY 10598, USA
Email: cneti@us.ibm.com

Gerasimos Potamianos
IBM T.J. Watson Research Center, Rte 134, Yorktown Heights, NY 10598, USA
Email: gpotam@us.ibm.com

Juergen Luettin
Robert Bosch GmbH, Automotive Electronics-Surround Sensing Systems, D-7152 Leonberg, Germany
Email: Juergen.Luettin@de.bosch.com

Eric Vatikiotis-Bateson
ATR International, 2-2-2 Hikaridai, Kyoto 619-0288, Japan
Email: bateson@isd.atr.co.jp

Computing is becoming increasingly ubiquitous and pervasive with multiple devices and modes of interaction. Traditional modes of human computer interaction based on a keyboard and a mouse are being replaced by more natural modes such as speech, touch, and gesture. This ongoing transformation towards pervasive and ubiquitous computing, induces the need for the next generation of human-computer interfaces (HCI) that are easy to use, transparent and robust in a variety of environments. It is important not only to improve the recognition and understanding of the individual modes (speech, gesture, etc.) in a variety of environments, but also to develop technologies and architectures that choose the most appropriate mode or modes to understand the content and the context of the interaction. Human cognition is an excellent guide to develop the next generation of HCI, since it depends on highly developed abilities to perceive, integrate, and interpret visual, auditory, and touch information. In spite of dramatically varying perceptual conditions, the human ability to organize and intelligently combine sensory data derived from multiple sensors, modulated by perceptual relevance and sensory confidence, is crucial for building a robust model of objects and events in our environment.

Although significant progress has been made in developing the individual modes of HCI based on audio (e.g., speech and speaker recognition), or visual input (e.g., face localization and person identification), only recently there has been an increased interest in exploiting the human capacity to process joint audio and visual information to improve systems for recognizing human activity (e.g., speech recognition, speech activity, speaker change, etc.), intent (e.g., speech intent), and identity (e.g., speaker recognition) in pervasive computing real-world environments. To realize the full benefit of joint processing in any of these application areas, several technical challenges have to be addressed, most notably the appropriate integration of multiple modalities. This special issue highlights innovative research in joint audio and video signal processing and its application to a variety of HCI critical areas, namely source localization, speech source separation in multispeaker environments, speech, and speaker recognition.

In the first paper, Zotkin et al. address the problem of joint audio-visual *source localization* using particle filters. They present a technique for tracking people by integrating audio cues captured using microphone arrays and visual cues captured using camera arrays. The multimodal tracking problem is formulated in terms of particle filters. The authors also show that the approach can be used for self-calibration, or to deal with situations where the sensors are moving, or where people are partially occluded.

The subsequent papers shift the focus to *audio-visual speech*. Of particular interest and important to developing HCI applications is the fact that visual information from the speaker's lower face both supplements and complements information provided by the traditional audio speech signal.

Sodoyer and his coauthors describe a novel approach to *separating speech sources* using the audio-visual coherence of the speech stimuli. In contrast to the classical blind source separation techniques, the authors explore the problem of

extracting a source from an additive mixture of speech sources. Using the visual lip tracking information of the desired speech source to extract its spectral envelope, the authors show that this approach compares favorably with independent component analysis.

Jiang et al. provide the first phonetically rigorous application of techniques that *correlate* speech behavior to various visual aspects of speech. They use multilinear techniques and measures of vocal tract articulator motion, face motion, and speech acoustics during production of sentences and nonsense syllables to characterize redundant properties of visible and audible speech behavior and the role of the vocal tract as a common source of phonetic information in both modalities. They show that the phonetic impact (e.g., of "place of articulation") on the degree of correspondence between audible and visible events is not uniform. The study also shows that speaker-specific differences in the strength of the cross-domain correlation in production do not necessarily match perceived differences in speaker intelligibility.

An important requirement for conducting research in joint audio-visual signal and speech processing is the availability of suitable *audio-visual speech corpora*. Patterson et al. describe an audio-visual speech database, CUAVE, that is compact in its organization and accessibility to other researchers and, at the same time, comprehensive enough to be useful in addressing a variety of methodological and research questions, especially those having to do with automatic detection and measurement of facial features and motion under varying position and orientation. The database consists of more than 7000 utterances (a total of 36 male and female speakers produced digits in isolation and in connected strings). In the latter part of the paper, the authors demonstrate the challenge and utility of the CUAVE database by extracting lip contours from moving and still faces and estimating a speaker-independent baseline for visual-only speech recognition, using various visual speech feature extraction techniques.

To allow joint audio-visual speech processing, reliable extraction of *visual speech features* is required. The problem consists of face detection, mouth region localization, and possibly lip (or, in general, face) contour estimation, followed by extraction of visual features that are informative about the uttered speech. These topics constitute the subject of the next cluster of papers in this special issue.

Daubias and Deléglise consider the first aspect of the visual speech processing problem; namely, they segment the lower mouth area into three classes of interest, that is, lip, inner mouth, and skin regions, by means of an artificial neural network (ANN) classifier (appearance based statistical model). The authors propose an automatic way of acquiring the labeled data needed to train the appearance model. Lip shape contours are estimated from aligned pairs of audio-visual sequences that contain the same phonetic context and whose alignment was calculated using dynamic time warping on the audio channel. The first pair sequences contain lips that are marked up with blue lipstick, and thus are easy to segment. The lip-shape model is built based on the easily extracted contours of such sequences. The second pair sequences contain unmarked lips, and thus, their segmentation is difficult, requiring both lip shape estimation and localization. The audio based alignment comes to the rescue, as it provides the lip-shape model parameters, thus reducing the problem to that of contour localization alone. The proposed method is reported to give excellent segmentation results, and saves the significant effort associated with data hand labeling.

Aleksic and his coauthors explore novel visual speech features based on MPEG-4 facial action parameters (FAPS). They describe an automatic method to extract the FAPS by combining active contour and template algorithms and demonstrate that these visual speech features can be used in conjunction with audio to improve medium vocabulary speech recognition in white Gaussian noise. Since MPEG-4 is a multimedia standard that deals with generic coding of audio-visual objects for multimedia applications, this work is an important step towards developing audio-visual speech recognition using standardized features.

Zhang et al. propose novel algorithms for both face segmentation and visual feature extraction. In particular, they employ Markov random fields for estimating the lip contours, and subsequently augment traditional lip contour based features with new visual features, that capture the tongue and teeth visibility. They demonstrate that the additional features help visual-only speech recognition. Finally, they report improved performance of both speech and speaker recognition on two standard audio-visual databases, by introducing the visual modality in addition to the traditional audio input.

The last paper of this cluster, by Gordan et al., concentrates on the issue of visual speech classification. Instead of employing the popular hidden Markov model (HMM) based recognizer, the authors propose a hybrid classification architecture that uses HMMs in conjunction with a parallel network of binary support vector machine (SVM) classifiers, the output of which provides posterior probabilities of the speech classes of interest (here, visemes). The SVMs operate on the pixel values of the mouth region of interest. Results are reported on a four-digit recognition task, and are compared to other visual feature extraction and classification methods.

Last but not least, one of the most exciting research topics in joint audio-visual speech processing is the subject of *integration (fusion)* of the two speech informative inputs. In this special issue, two papers concentrate on this topic.

Heckmann et al. use a hybrid HMM/ANN architecture for audio-visual speech recognition, and integrate the audio and visual classifiers at a likelihood, decision level, in a "state synchronous" fashion. They first consider various schemes to weigh the posterior likelihoods of the audio and visual-only ANNs with appropriate stream exponents ("weights"), and they conclude that their multiplicative combination, which respects their class-conditional independence, is superior. Subsequently, they concentrate on the adaptive estimation of the combination weights, based on the reliability of each stream of information, as captured by three criteria. Results are reported on a single-speaker, connected digits database.

Nefian et al. take a different approach and concentrate on "state-asynchronous" architectures for audio-visual fusion, by means of HMMs. They consider integration both at the feature, as well as at the likelihood (decision) level, using a multitude of Bayesian network models, such as the product HMM, the factorial HMM, and the coupled HMM. They present iterative algorithms for obtaining maximum likelihood estimates of the model parameters, as well as their initial estimates. They also compare the complexity of the models, both in terms of number of parameters and recognition time. Finally, they report their performance on a speaker-independent audio-visual speech recognition task of isolated words.

As a final note, we would like to express our thanks to all the contributing authors and reviewers for this special issue. We believe that this is a very exciting area of research, and hope that these papers will further stimulate work on joint audio-visual signal and speech processing.

Chalapathy Neti
Gerasimos Potamianos
Juergen Luettin
Eric Vatikiotis-Bateson

Chalapathy Neti received his B.S. degree in electrical engineering from the Indian Institute of Technology, Kanpur, and his M.S. degree in electrical engineering from Washington State University, and his Ph.D. degree in biomedical engineering from The Johns Hopkins University. He is a Research Manager in the Human Language Technologies Department at IBM T.J. Watson Research Center in Yorktown Heights, NY. He currently manages a research effort on audio-visual speech technologies. In this role, he is leading a group of researchers in developing algorithms and technologies for joint use of audio and visual information for robust speech recognition, speaker recognition, and multimedia content analysis and mining. He has been with IBM since 1990. Dr. Neti's main research interests are in the area of perceptual computing (using a variety of sensory information sources to recognize humans, their activity and intent), speech recognition, multimodal conversational systems for information interaction and multimedia content representation for search and retrieval. He has authored over 30 articles in these fields and has five patents and several pending. He is an active member of IEEE, a member of IEEE Multimedia Signal Processing technical committee, and an Associate Editor of IEEE Transactions on Multimedia.

Gerasimos Potamianos graduated in electrical and computer engineering from the National Technical University of Athens, Greece, in 1988, and received the M.S.E. and Ph.D. degrees in electrical and computer engineering from the Johns Hopkins University, Baltimore, Maryland, in 1990 and 1994, respectively. From 1994 to 1996, he was a Postdoctoral Fellow with the Center for Language and Speech Processing, Baltimore, Maryland, working on decision tree based language models and smoothing techniques. From 1996 till 1999, he was a Senior Member of Technical Staff with the Speech and Image Processing Services Laboratory at AT&T Labs-Research, working on robust techniques for audio-visual automatic speech recognition. In September 1999, he joined the Human Language Technologies group at the IBM T.J. Watson Research Center as a Research Staff Member, currently continuing his work on audio-visual speech recognition, within the Audio-Visual Speech Technologies group. His research interests span the areas of signal and image processing, multimedia signal processing, automatic speech recognition, language modeling, audio-visual speech processing, recognition, and synthesis, by means of statistical methods. He has published over 30 articles in these areas and has four patents filed. He is a member of IEEE and a member of the Technical Chamber of Greece.

Juergen Luettin studied electrical engineering in Karlsruhe, Nottingham, and Sheffield and received the Ph.D. degree from the University of Sheffield, UK in 1997 on the subject of visual speech and speaker recognition. From 1996 to 1997 he was a Research Assistant at IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence) in Martigny, Switzerland where he worked on speaker verification, sensor fusion, and speech recognition. From 1997 to 2000 he headed the computer vision group at IDIAP and contributed to research topics in biometrics, speech recognition, face analysis, handwriting recognition, and video retrieval. He has also initiated and participated in several EU IST projects as well as Swiss SNF projects. From 2000 to 2002, he led the pattern recognition group at Ascom AR&T, Switzerland, where he was responsible for software and product development in image and speech processing, with the main focus on intelligent video surveillance. In 2002, he joined the Automotive Electronics Division of Robert Bosch GmbH, Germany. Dr. Luettin has been an invited scientist at Johns Hopkins University in 1997 and in 2000, served as a reviewer of several international scientific journals and has authored or coauthored over 40 scientific articles.

Eric Vatikiotis-Bateson was born in Washington, DC in 1952. He received a Bachelor's degree in philosophy and physics from St. John's College, Maryland, in 1974. He received a certificate in ethnographic film making in 1976, and an M.A. in Linguistics from Indiana University in 1978. From 1982 to 1987, he was an NIH pre-doctoral fellow at Haskins Laboratories (Connecticut) investigating "the organization and control of speech production." He received a Ph.D. from Indiana University in 1987 and was appointed Staff Scientist at Haskins Labs in 1988. He has been at ATR International in Japan since 1990, while maintaining affiliations with Haskins Laboratories (USA), the Institute for Phonetics and Spoken Communication at Munich University (Germany), and the Institute for Spoken Communication (ICP, France). Since 2000, he has headed the Communication Dynamics Project around which Department 2 of the newly-formed Human Information Science Lab was formed in 2001. The goal of this project is to examine the structure and processing of communicative events in complex auditory and visual environments. From January, 2003 he will hold a chair in Speech and Cognitive Science at the University of British Columbia in Vancouver, Canada.