

Audio-Visual Technologies for Lecture and Meeting Analysis inside Smart Rooms

Gerasimos Potamianos

Human Language Technologies Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
Email: gpotam@us.ibm.com

Abstract

Analysis of lecture meetings recorded inside smart rooms has recently attracted much interest, being the focus of international projects and technology evaluations. In this keynote, we briefly overview one such project, “Computers in the Human Interaction Loop” (CHIL), with emphasis on the perceptual technology components developed. In particular, we focus on person tracking and speech processing technologies, and present the developed IBM systems.

Keywords: Speech Processing; Speech Recognition; Speaker Diarization, Speech Activity Detection; Visual Tracking; Face Detection; Meeting Data.

Brief Overview of the Keynote

Interactive lectures and meetings play significant role in human collaborative activities in the workplace. Not surprisingly, analysis of interaction in these domains has attracted significant interest in the community, being the focus of a number of research efforts and international projects, for example CHIL, AMI, and the U.S. National Institute of Standards and Technology (NIST) Smartspace effort. In these projects, the interaction happens inside smart rooms equipped with multiple audio and visual sensors. Based on the resulting captured data, the goal is to extract higher-level information in order to assist, for example, in lecture meeting indexing, browsing, summarization, and understanding. To achieve this, technology components need to be developed that address basic questions about the “who”, “where”, “what”, and “when” of the interaction.

Addressing these goals is the main aim of the CHIL EU-funded project, run under the technical coordination of the Interactive Systems Laboratories at the University of Karlsruhe, Germany (CHIL project website). In CHIL, computers are reduced to “discreet” observers of human activity through the use of far-field sensors, and are to provide lecture meeting support services to the participants, based on a common architecture that integrates perceptual components. Central to this goal are people tracking and speech processing technologies; in particular *face detection* and three-dimensional (3D) *person tracking*, as well as automatic speech recognition (ASR) or *speech-to-text* (STT), and its complementary technologies, *speech activity detection* (SAD) and *speaker diarization* (SPKR). Significant research effort is be-

ing devoted to developing robust and efficient algorithms to attack these problems. Noticeably, these efforts have been rigorously evaluated in the past few years through project-internal evaluation campaigns, the NIST-sponsored Rich Transcription (RT) Meeting Recognition Evaluation (RT06s evaluation website), and the recent CLEAR (Classification of Events, Activities, and Relationships) campaign (Stiefelhagen et al. 2006).

In this talk, we present a summary of the IBM efforts with respect to the CHIL project, with emphasis on the developed technologies to address face detection, 3D tracking, SAD, SPKR, and STT for the lecture meeting scenario, central to CHIL. Both vision and speech tasks are particular challenging: Speech-wise, due to the presence of multiple speakers with often overlapping speech, a variety of interfering acoustic events (chairs moving, door noise, typing, computer noise, etc.), the strong accents of most speakers and interacting audience members, a high level of spontaneity, hesitations and disfluencies, the technical seminar contents, the relatively small amount of in-domain data, and the use of far-field sensors (Huang et al. 2006); vision-wise, due to low-resolution distant data, people occlusion, and lighting variations (Potamianos and Zhang 2006). Nevertheless, the work reported here shows that addressing these problems in real human interaction scenarios such as CHIL lecture meetings is achievable.

Acknowledgements

A number of colleagues at IBM have been instrumental to the presented work: Stephen M. Chu, Stanley Chen, Jan Curin, Pascal Fleury, Jing Huang, Brian Kingsbury, Jan Kleindienst, Vit Libal, Etienne Marcheret, Chalapathy Neti, Daniel Povey, Thomas Ross, Larry Sansone, Andrew W. Senior, Roberto Sicconi, Olivier Siohan, Alvaro Soneiro, Martin Westphal. Furthermore, Amrith Tyagi and Zhenqiu Zhang have worked on the topics discussed in this work during summer internships at IBM. Support for this work by the European Commission under integrated project CHIL, “Computers in the Human Interaction Loop”, is also acknowledged.

References

- The CHIL Consortium Website, <http://chil.server.de>
- Rich Transcription 2006 Spring Meeting Recognition Eval., <http://www.nist.gov/speech/tests/rt/rt2006/spring>
- Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., & Soundararajan, P. (2006), “The CLEAR 2006 evaluation,” *Lecture Notes in Computer Science*, Stiefelhagen, R. & Garofolo, J. (eds.), Vol. 4122 (In Press).
- Potamianos, G. & Zhang, Z. (2006), “A joint system for single-person 2D-face and 3D-head tracking in CHIL seminars,” *Lecture Notes in Computer Science*, Stiefelhagen, R. & Garofolo, J. (eds.), Vol. 4122 (In Press).
- Huang, J., Westphal, M., Chen, S., Siohan, O., Povey, D., Libal, V., Soneiro, A., Schulz, H., Ross, T., and Potamianos, G. (2006), “The IBM Rich Transcription Spring 2006 Speech-To-Text System for Lecture Meetings,” *Machine Learning for Multimodal Interaction* (In Press).