A System Structure for Multimodal Emotion Recognition in Meeting Environments

Ronald Müller, Sascha Schreiber, Björn Schuller, Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Arcisstrasse 16, 80333 München, Germany
{mueller,schreiber,schuller,rigoll}@mmk.ei.tum.de

Introduction

In this contribution we introduce a system structure capable for robust multimodal Emotion Recognition in Meetings as it is proposed to be applied within the EU-IST Integrated Project AMI (Augmented Multiparty Interaction) [2]. Analysis of human affects is a must on the way to a full understanding of meeting participants' communication. Exemplarily, issues addressed by Emotion Recognition tasks in meetings are the agreement or disappointment of people on decisions made, the curiosity or disinterest of participants on specific topics, or the question on the general social mood, meetings are held in at large companies, to allow for suggestions about the overall employee satisfaction.

Research on Emotion Recognition in AMI needs to adapt on the technical setup of the meeting room. The proposed setup, designed for up to six participants, comprises three wide-angle cameras, microphone arrays in the middle of the conference table, and -most relevant for speech and facial Emotion Recognition- lapel microphones as well as close-up cameras, positioned on the table in front of each seat.

System Structure and Modules

Figure 1 shows the proposed structure of a system for multimodal Emotion Recognition in meeting environments, as it is addressed within AMI project. In the proposed poster the relevant system modules shall be introduced with respect to their requirements due to the task conditions, their inputs, outputs as well as the underlying algorithms. The technical implementation is partly accomplished, partly subject of current or upcoming research activities.

The audio-visual capture comprises synchronized streams from the close-up camera and the lapel microphone of one meeting participant. The strived system taps and fuses 5 information sources derived from the multimedial data stream, namely the semantic content of speech, the sound of speech, facial expressions, body pose, and hand gestures, while our current research focuses on the first three mentioned. Information about previous works can be found in [1] [3] [4] [5]. Most interest will lie on the Multilayered Multimodal Fusion, which needs

From: MLMI 2005, Edinburgh, UK

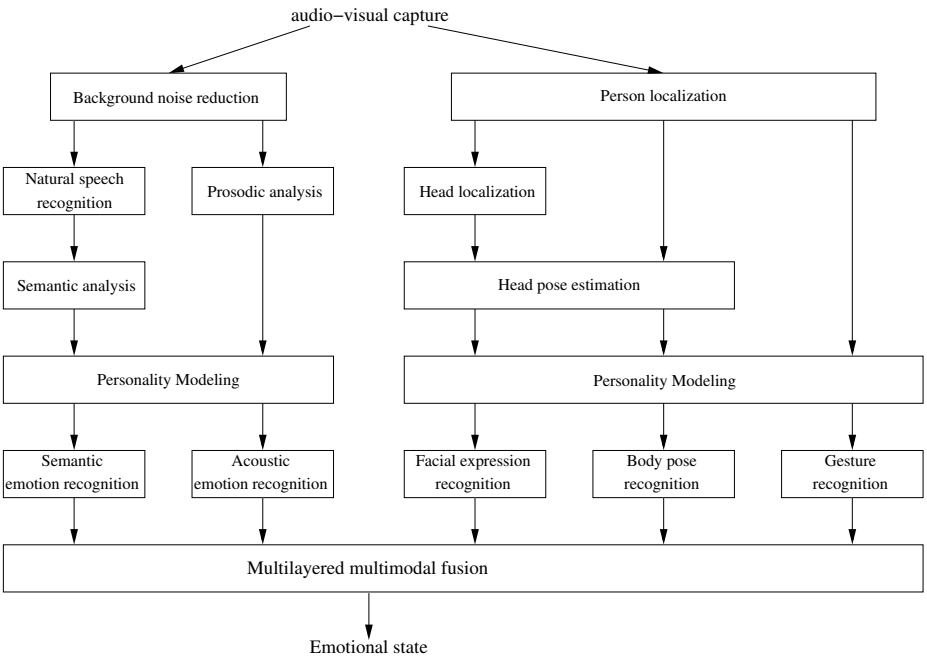

Fig. 1. System structure for multimodal Emotion Recognition in meeting environments

to cope with uncertain recognition results, the drop out of information sources, variable reliabilities of single streams and estimations belonging to different video segments of different time length.

Acknowledgment This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (FP6-506811, publication).

References

[1] R. Müller, B. Schuller, and G. Rigoll. Enhanced robustness in speech emotion recognition combining acoustic and semantic analyses. In *Proc. of Workshop From Signal To Signs of Emotion and Vice Versa*, Santorin, Greece, 2004. EU-IST Network of Excellence Humaine.

[2] EU-IST Project AMI (Augmented Multi party Interaction) www.amiproject.org.

[3] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recogntiion combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture. In *Proc. of ICASSP 2004*, Montreal, Canada, 2004. IEEE.

[4] B. Schuller, R. Villar, and M. Lang G. Rigoll. Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In *Proceedings of ICASSP 2005*, Philadelphia, USA, March 2005. IEEE.

[5] S.Schreiber and G.Rigoll. Robust face tracking and person action recognition in mettings. In *MLMI 2004*, Martingny, Switzerland, June 2004.