On the Possible Role of Acoustics for Multimodal Analysis and Recognition of Human Behavior in Smart Environments

Gerhard Rigoll

Lehrstuhl für Mensch-Maschine-Kommunikation, TU München D-80333 München, Germany
Email: rigoll@ei.tum.de

Introduction

The analysis and recognition of human behavior and action in smart environments is a new challenging discipline in the area of multimodal interfaces. Smart environments are environments that are capable of interacting with the user or supporting the actions of the user. Examples for such anvironments are e.g. Smart Homes, Smart Offices, or Smart Meeting Rooms. In all those environments, a new trend in human-machine-communication is to enable the user to interact with the devices and the equipment of the room. For that purpose, the smart environment has to be aware of the presence of the user(s) and to be able to detect their current actions. Also verbal actions of the user are crucial for detecting his intentions, and therefore, speech recognition in rooms is an important discipline for such a research purpose, which automatically involves the use of room acoustics for recognition. Room acoustics will also play a major role in audio-visual tracking of the users, since their position has to be always known. In this paper, some examples for the possible use of acoustics will be presented in the framework of a recently started large EU-funded integrated project on smart meeting rooms.

Smart Meeting Room Equipment

Smart meeting rooms can be seen as part of a relatively new research field that could be called "smart environments" and rapidly gain more and more attention in the broad area of human-machine communication. Other sub-areas of smart environments are for instance "smart living rooms", "smart offices" or even "smart parking lots". All of these environments have the common feature that they are equipped with a - possibly very large number – of sensors that enable the environment to communicate with the humans that are in this environment. This communication ability makes such environments "smart" in that sense, that the user can interact with devices or parts of the room and the environment is able to assist the user in his actions or intentions. In that way, the environment itself becomes interactive and the user typically performs some multimodal interaction in that environment which does not only interact with the user but also assists the user and records the events that take place while the user interacts in such an environment. In smart meeting rooms, the sensors consist of microphones and cameras for event recording. Fig. 1 shows an outline of the smart meeting room as it is currently realized and used in the EU-sponsored project M4 (see [1]) by one of M4 project partners. It can be seen that this one is equipped with individual microphones for each speaker and several microphone arrays. There are also several cameras installed in such an order, that each camera can capture three meeting participants. An additional wide angle camera is installed to capture a complete view of the entire meeting scenario.

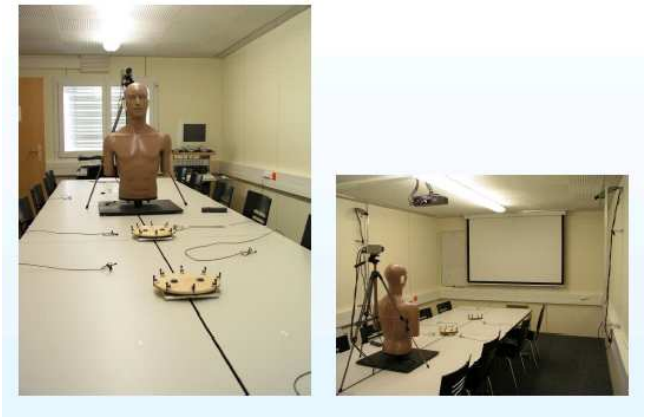

Fig. 1: Example for a smart meeting room

Clearly, the idea is to record the audiovisual data from the sensors described above and to analyze this data in a later off-line stage in order to transcribe as precisely as possible what happened in that meeting. This concerns of course the transcription of the spoken utterances in the meeting, however is not restricted to audio data, but also takes into account the captured video tracks in order to analyze what actions have been performed by the meeting participants. The final goal is the use these transcriptions in a meeting browser, that basically contains the minutes of the meeting in a multimedia format, containing audio and video. With those transcriptions, it is then possible to query in this browser e.g. for the discussions of the participants or to search who gave a presentation and if some voting took place. Such an information will be extremely valuable e.g. for people who could not participate in the meeting or for participants who want to see the summary of the meeting.

Speech and Audio Processing

It would be beyond the scope of this paper to describe all the audio processing techniques that have to be deployed in meeting transcriptions. It is obviously clear that speech transcription in meetings is an extremely demanding speech recognition task, due to the following facts: Speech of meeting participants in such environments is a typical case for spontaneous speech recognition, since in meetings, almost every utterance is spontaneous, with hesitations, repetitions, stuttering and many other effects, except perhaps well-prepared meeting presentations, which could be closer to read speech. However, especially in meetings there are additional burdens, such as e.g. cross-talking, background talking, background noise or far distance microphones.

As in many of today's speech recognition problems, an appropriate database is also in this project one of the major

key factors for success. Since there are almost no databases for meeting recordings available – and especially not publicly available – a special M4 database is currently recorded. Additionally, the consortium has access to the ICSI meeting corpus (see [3]), where 75 meetings (72 hours) of fully annotated meeting recordings have been collected, however under different acoustic conditions as foreseen in the M4 project.

Speech recognition experiments for the M4 environment are under way, where in a first stage, training is performed on the ICSI data and the Switchboard database, which is partially suitable due to the fact that it also contains spontaneous speech. A recognition error rate of around 30-40% (WER) is expected to be obtained with this approach.

Multimodal Processing of Audiovisual Data

As already mentioned, in meetings, other communication channels than audio alone are valuable sources of information. This is especially true for the visual information channel, that can be evaluated in order to analyze the events and actions that were performed during a meeting. Very often, these visual cues are in strong correlation with acoustic cues, e.g. when a meeting participant gives a presentation in front of the other participants, draws a sketch on the board and explains his presentation by spoken comments. If such information is intended to be evaluated and interpreted, the key for this is to find the position of the meeting participants. If this is feasible, then it is possible to find out who did something, at what time it was done and what has been done by this person. In meetings, the most typical position of people is when they are sitting around the meeting table and with sufficiently high resolution of the camera, the most efficient way to locate the position of a person is to localize his face. Neural network techniques in conjunction with colour analysis can be successfully employed for this task. Action recognition is another important issue for meeting transcription. In this case, a few pre-defined actions are described, such as e.g. entering, leaving, rising, nodding or voting. Special measures are taken and described in more detail in [2] to extract global motion features from the recordings of such actions and to classify them with Hidden-Markov-Models. A respectable action recognition rate of up to approx. 80% can be reached. The further reaching goal is the multimodal recognition of audio-visual data streams, in order to recognize events such as presentations, group discussions, or even meeting interruptions, such as e.g. coffee breaks.

Meeting Browser

Finally, the expected result of the meeting transcription process is the display of the transcription results in a so-called meeting browser. Fig. 2 shows an example for such a system. The basic idea is that – similarly to browsing through websites – it should be possible to load the transcription and index files resulting from the multimodal meeting analysis and indexing into the browser and to navigate through these results and query the system e.g. about the content of the speech transcription, the tracking of various speakers, the analysis of the previously discussed talkativity of speakers, or the search for complex events, such as votings or presentations of participants.

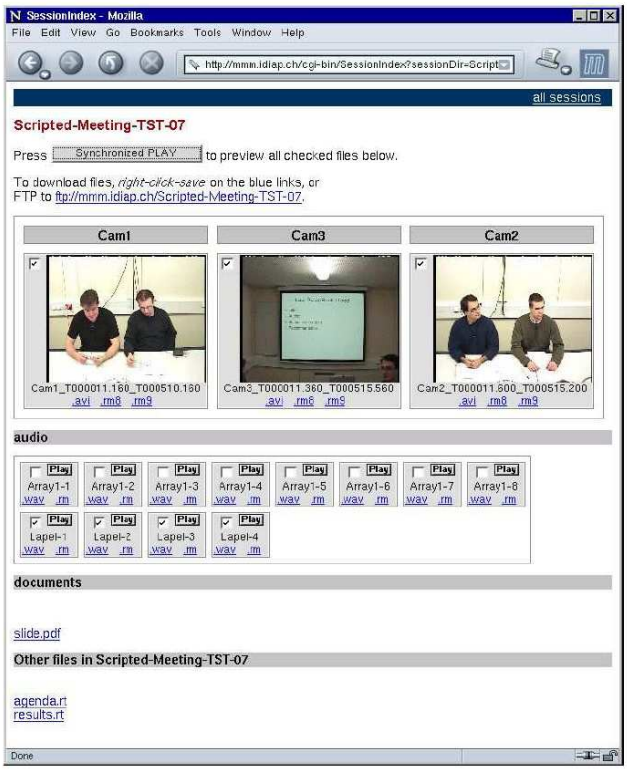

Fig. 2: Example for a multimodal meeting browser

Summary and Conclusion

The purpose of this paper was to introduce the research area of smart meeting rooms as new field of research that is also interesting for applied research in acoustics. Possible challenges for acoustics include:

- ✓ far distance microphone speech recognition
- ✓ source localization for speaker separation
- ✓ tracking of static & moving sound sources using binaural auditory models
- ✓ multimodal fusion for audio-visual person tracking

It is expected that the area of meeting analysis and smart meeting rooms will have an enormous development in the next years to come and much further sophisticated methods will be required and developed in order to make the vision of automatic meeting transcription and browsing a reality.

References

[1] http://www.dcs.shef.ac.uk/spandh/projects/m4/

[2] Zobl, M., Wallhoff, F. and Rigoll, G.: Action Recognition in Meeting Scenarios Using Global Motion Features. *Proc. IEEE Workshop on Performance Evaluation of Tracking and Surveillance, Graz, Austria, March 2003.*

[3] Janin, A, Baron, D, Edwards, J, Ellis, D, Gelbart, D, Morgan, N, Peskin, B, Pfau, T, Shriberg, E, Stolcke, A and Wooters, C. The ICSI Meeting Corpus. *Proc. ICASSP-03, Hong Kong, April 2003.*