SMaRT: THE SMART MEETING ROOM TASK AT ISL

Alex Waibel, Tanja Schultz, Michael Bett, Matthias Denecke, Robert Malkin,
Ivica Rogina, Rainer Stiefelhagen, and Jie Yang

Interactive Systems Laboratories (ISL), Carnegie Mellon University and Karlsruhe University Germany

ABSTRACT

As computational and communications systems become increasingly smaller, faster, more powerful, and more integrated, the goal of interactive, integrated meeting support rooms is slowly becoming reality. It is already possible, for instance, to rapidly locate task-related information during a meeting, filter it, and share it with remote users. Unfortunately, the technologies that provide such capabilities are as obstructive as they are useful – they force humans to focus on the tool rather than the task. Thus the veneer of utility often hides the true costs of use, which are longer, less focused human interactions. To address this issue, we present our current research efforts towards SMaRT: the Smart Meeting Room Task. The goal of SMaRT is to provide meeting support services that do not require explicit human-computer interaction. Instead, by monitoring the activities in the meeting room using both video and audio analysis, the room will be able to react appropriately to users' needs and allow the users to focus on their own goals.

1 INTRODUCTION

Computing and communication are merging and therefore providing an increasingly integrated and interactive environment which could substantially support humans in various tasks and situations. Yet, it is a common experience for most of us, that the technology that provides so many benefits, appears to become a burdensome chore, and intrusion in our lives. To a large extent this intrusion stems from the fact that machines demand our undivided attention and force us to careful monitor the technology to receive and select the services we seek at any given moment. Advances have certainly been made to improve this situation by creating more intuitive interfaces and by extending the interaction between humans and computers to include other modalities, like speech, pointing, handwriting, and dialogs between human and machine. While this has improved the resulting interfaces considerably, the overall human experience with computers, however, is still quite unsatisfactory: Rather than control, today's interactive devices have generated added confusion, irritating distractions, information overload and preoccupation with the technology instead of freeing the user from it to work the *human* problems at hand. In order to improve on this problematic world, we aim at creating a new human workspace in which human interaction remains focused on the interaction *with other humans* and on actual human activities. To achieve this goal the computer must be equipped with perceptive capabilities to capture the relevant information about the humans' needs and the context in which they act. The computer is expected to intrude as little as possible and must also learn from its interaction with the environment and people.

In this paper we present our current research efforts towards "SmaRT: the Smart Meeting Room Task" at Interactive Systems Laboratories. The goal of our work is to provide a smart meeting room that supports humans in any kind of meeting situations. The functionality of such a smart meeting room stretches from *pre-meeting* activities like calling meeting participants, and navigating them to the meeting room, creating a pleasant room atmosphere (switch cell phones off, adjust lighting conditions, regulate heating system, brew coffee, etc.) to *in-meeting* services like support presentations (automatically switch presentation slides, provide information on demand over web) and fully automatically create notes and summaries of the meeting, and further to *post-meeting* activities like archiving and indexing meeting summaries as well as audio and video recordings to make them retrievable at a later state.

Figure 1: SmaRT data collection

2 MOTIVATING EXAMPLE

Imagine a meeting room equipped with smart room technology, and what might happen in such a room when a meeting takes place. Before the meeting, the room, sensing no human presence, is running it's self-cleaning processes. This activity is interrupted when the acoustic scene analysis component recognizes the sound of a key in the door. As several humans wander into the room, the vision systems identify them and search for their personal websites for possible later use. Since the humans are still wandering around the room and chatting, the room decides that a meeting is imminent but has not yet started; the meeting room's coffee subsystem is activated and the room begins to play soft music. Finally, the participants sit at the conference table and the meeting begins. The music stops, and as the participants introduce themselves, the cached web pages appear on a display. In the background, automatic minute-taking procedures have started. The chair of the meeting outlines the agenda, and the minute-taking software builds a graphical representation of the agenda. The meeting proceeds, with the

room taking notes of who is speaking, who is being spoken to, and what topics are being covered. During the meeting, the telephone rings. The acoustic scene analysis system identifies the event, and the room notes that the call is coming from a participant who has not yet shown up for the meeting. The person who answers the phone mentions that he will send the current meeting status to the absent participant; the meeting room, using key phrase spotting, interprets this statement as a request and sends the update. Shortly afterwards, the absent participant arrives; unfortunately, she does not have clearance to view the materials currently being displayed on the projection screen. The vision system recognizes her, and another process, noting the lack of appropriate permissions, blanks out the projection screen. The meeting chair decides that the new arrival should be allowed to see the material, and verbally requests a clearance change. Finally, the meeting draws to a close. The chair, aided by the meeting minute system, summarizes the action items for all present and publishes the minutes on a secure website for later perusal. As the participants begin to rise and file out, both the acoustic and visual systems note the change; as the last humans leave, the room returns to its original state.

3 HUMAN INTERACTION

In order to provide the above mentioned functionality, the most important prerequisite is the interpretation of human (inter)actions and communication. The computer needs to know what humans do, how and with what or whom they interact or to what they refer. The type of questions which needs to be asked are: Who is speaking?, What is spoken?, How is it said?, and Where or to Whom it is addressed?. To answer to these questions the computer needs to recognize and understand multi-modal cues in human-to-human inter-actions. To answer the question about *Who is speaking?* we use people identification, speaker identification, and face identification. It also includes gathering information about the type of person (dominant, submissive, etc.) and the relationship between communication partners. To answer the question *What is spoken?*, we include speech and discourse modeling, i.e. speech recognition, lip reading, discourse states (speech acts, topics), turn taking, as well as discourse types and genres (negotiation, chatting, lecturing, etc.). Important information about a meeting are also revealed by the fact *How?* something is said. This requires the classification of the emotional state of a person (happy, sad, afraid), as well as the status of excitement (busy, nervous, relaxed, tired) as well as the discourse style of the conversation (sloppy, formal, colloquial). Finally, the *Where or to Whom?* needs to be answered by speaker localization as well as the determination of the focus of attention of a speaker. In the remainder of the paper we describe how we use verbal (words, speakers, emotion, language, summaries, topics, handwriting) and visual (identity, gestures, body-language, face, gaze, pose, facial expressions, focus of attention) towards the creation of such a smart meeting room environment.

The resulting smart meeting room systems should reduce workload in measurable ways. To achieve this breakthroughs in a number of component technologies, the integrated system and a better understanding of its new use in human spaces are needed. Evaluation must be carried out both in terms of performance and effectiveness to assess and track progress. In order to enable the design, development, and evaluation we started to collect data in various scenarios, from many speakers, in several languages and accents, under different conditions [4].

4 INFORMATION/ACTIVITY AWARENESS

Many semantically relevant events and state changes in the meeting room are accompanied by auditory cues. These cues can include ringing telephones, knocks on doors, doors opening and closing, footsteps, or even the subtle sound texture difference between many people speaking informally and a single person leading a discussion. These events and state changes are used to trigger a number of responses by the smart room; for example, shifting to pre-meeting mode when the door opens or shifting to meeting-interrupted mode when the telephone rings. To capture these events and state changes, we use a simple machine listening system trained to recognize these events and states. This system uses a set of discrete HMMs which continually monitors the incoming audio signal to detect known events. Experimental results for this system are forthcoming.

5 PEOPLE TRACKING AND IDENTIFICATION

Tracking people is a common problem for developing an intelligent space. We have been developing robust people tracking technologies for an indoor environment. These technologies enable to track people from a single camera, multiple cameras, and an omnidirectional camera [6,9,10]. Figure 2 illustrates a system for real-time segmentation and tracking of individuals moving across realistic backgrounds using multiple cues. Different windows in Figure 2 in turn, show background, foreground, motion segmentation, color segmentation, and joint segmentation. The segmentation and extraction algorithm exists in multiple camera configurations, trading speed for accuracy according to the demands of real-time applications. We have further developed technologies for people identification using multiple cues such as face, voice, and color appearance [10,11]. From our experience, color appearance is one of the most robust features for tracking in a complex scene such as a meeting room.

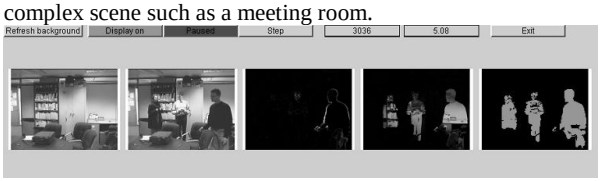

Figure2: Real-time segmentation and people tracking

6 SPEAKER, ACCENT, AND LANGUAGE IDENTIFICATION

For the identification of non-verbal cues from spoken speech, namely speaker, accent, and language, we recently presented a joint framework which uses phone strings, derived from different phone recognizers, as intermediate features and performs classification decisions based on their perplexities [3]. By using information derived from phonotactics, we expect to cover speaker idiosyncrasy and accent-specific pronunciations. We also anticipate greater robustness under mismatched conditions since the information is provided from complementary phone recognizers. Our identification results validate this concept. The evaluation on our distant microphone database proved the robustness of the approach, achieving a 96.7% speaker identification rate on 10 seconds of audio under mismatched conditions, clearly outperforming Gaussian Mixture Models on large distances. Furthermore we achieved 97.7% accent discrimination accuracy between native

and non-native English speakers. For language identification, we obtained 95.5% classification accuracy for utterances 5 seconds in length and up to 99.89% on longer utterances. The speaker and accent identification experiments were carried out on English data, although none of the applied phone recognizers were trained or adapted to English spoken speech. Similarly, our language identification experiments were run on languages not presented to the phone recognizers for training. The language independent nature of our experiments suggests that they could be successfully ported to non-verbal cue classification in other languages [3].

7 MEETING BROWSER

An important part of meeting recognition is the ability to efficiently capture, manipulate and review all aspects of a meeting. To that end we have developed a meeting browser [7,8] that lets users: (1) Create meeting records and transcriptions of meetings with participants remotely located; (2) Create and customize dialogue, audio, and video summaries to the user's particular needs; (3) Create a database of corporate knowledge. (4) Quickly and accurately create and disseminate a list of conclusions and action items; (5) Provide rapid access to meeting records to allow browsing and reviewing existing meetings; and (6) identify for each utterance the speaker properties (type, social relationships, and emotion) as well as the discourse structure and type.

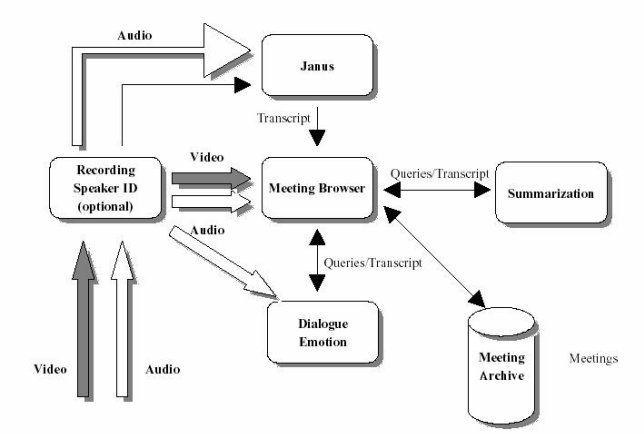

Figure 3: Component of the Meeting Room System

When a meeting is being created, each participant may join either remotely or locally. Once the meeting has begun, speech is transmitted to JANUS, our speech recognition engine. As the speech is recognized, the hypothesis is sent to the dialogue system where it is assembled into a meeting format. The meeting browser displays the transcript for the current meeting. The meeting transcript can be sent to a summarization system which will create a summary of the current dialogue. Finally, a user may elect to save a meeting including any summaries in the meeting archive from within the meeting browser. At the end of meetings, it is customary to reiterate a set of action items. Using speech recognition, we recognize the items and mail them out to each of the meeting participants. Likewise, we can mail complete meetings, meeting segments, or summaries including the audio portion directly from within the meeting browser to meeting participants or any other interested parties. Each of these may include annotations, comments or corrections. Corrections can be done by using a keyboard or handwriting recognition using a handwriting recognizer developed in our lab.

8 LECTURE TRACKER

Archiving, indexing, and later browsing through stored presentations and lectures is a task that can be observed with a growing frequency. We have investigated the special problems and advantages of lectures and developed a mechanism for adapting a speech recognizer towards a lecture such that the recognition accuracy can be significantly improved by prior analysis of the presented documents using a special class-based language model. We extract important words from the documents, retrieve linguistic information about them through internet search engines, and add them, even if they are not OOV, into one or more carefully selected language model classes. This approach led to a reduction of the word error rate by up to 20%. We have defined a tracking accuracy measure which measures how well a system can automatically align recognized words with parts of a presentation (i.e. the portion of the time in which the system predicts the correct slide for automatic slide switching). We were able to reduce the tracking error by up to 18% by prior exploitation of the presented documents [1].

9 FOCUS OF ATTENTION

We have also addressed the problem of tracking the focus of attention of participants in a meeting, i.e. we try to detect who is looking at what or whom during a meeting [5]. Such information can for example be used to control interaction with the meeting room or to index and analyze multimedia meeting records.

Figure 4: Views of meeting participants, captured with an omnidirectional camera.

In the developed system participants are simultaneously tracked in a panoramic view and their head poses are estimated using neural networks. For each participant, probability distributions of looking towards other participants are estimated from their head orientations using an unsupervised learning approach. These distributions are then used to predict focus of attention given a head pose. The accuracy of such predication is 73% accurate in detecting the participants' focus of attention on our test data. Furthermore, we have investigated how focus of attention can be predicted based on knowledge of who is currently speaking, and how this audio-based prediction can be improved by taking the history of utterances into account. On the recorded meetings, participants' focus of attention has been predicted correctly in 63% of the frames by using audio information only. In addition, we have shown how the audio- and the video-based predictions can be fused to get a more accurate and robust estimation of participants' focus of attention. By using both head pose and sound, focus of attention could be detected in 76 % of the frames in recorded meetings. To answer how precisely focus of attention can be predicted in a meeting just based on the participants' head orientations we have recorded and analyzed eye gaze and head orientations of four subjects in a meeting [2]. The user study

clearly demonstrated that head orientation is a reliable cue to detect at whom someone is attending to. In the meetings which we recorded for this study, we were able to correctly determine at whom the subject was looking in 89% of the time just based on the subject's head orientation. Finally, we have investigated how a neural network for head pan estimation can be adapted to work in a new location. Our experiments showed that adaptation images from only four subjects were sufficient to achieve good focus of attention detection accuracy in a new location with completely different illumination conditions.

10 SYSTEM ARCHITECTURE

In order to integrate these disparate components into a single smart room system, we have created a flexible, robust smart room architecture designed to work with arbitrary components. This architecture, shown in figure 5, is made up of five main components: the Controller, the Interaction Model, the Message Listener, and the Action Dispatcher. The Message Listener accepts information from the Smart Room subsystems, and forwards messages to the Controller. The Controller then checks the Interaction Model, which is a finite state machine containing conditions and actions for each state. If the incoming message matches a condition in the current state of the Interaction Model, an Action – which is just some command that one of the Smart Room subsystems can carry out – is generated and sent to the appropriate subsystem via the Action Dispatcher. One special type of Action is the „change state" action, which is performed by the Controller. This architecture allows for Smart Room subsystems to interact in interesting and useful ways while maintaining complete modularity. The only steps needed to import a new subsystem are to add the appropriate conditions and actions to the Interaction Model before system startup and register the new subsystem with the Controller at runtime.

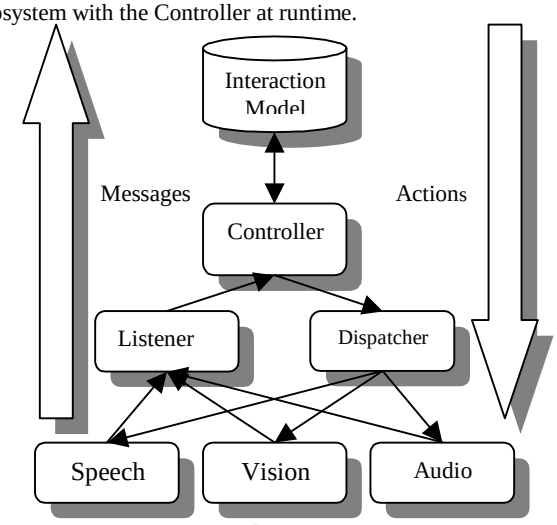

Figure 5: SmaRT system architecture

CONCLUSIONS

In this paper we presented our current research efforts towards SMaRT: the Smart Meeting Room Task. The goal of SMaRT is to provide meeting support services that do not require explicit human-computer interaction. Instead, by monitoring the activities in the meeting room using both video and audio analysis, the system is able to react appropriately to users' needs and allow the users to focus on their own goals. It supports human-machine, human-human, and human-computer-human interactions providing multimodal and fleximodal interfaces for multilingual, multicultural meetings.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge all members of the Interactive Systems Labs which helped to develop the system and to collect the data.

REFERENCES

[1] I. Rogina and T. Schaaf: Lecture and Presentation Tracking in an Intelligent Meeting Room. Proceedings of the ICMI 2002, Pittsburgh, PA, October 2002.

[2] R. Stiefelhagen and J. Zhu: Head Orientation and Gaze Direction in Meetings. Conference on Human Factors in Computing Systems (CHI2002), Minneapolis, April, 2002.

[3] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel: Speaker, Accent, and Language Identification using Multilingual Phone Strings. Proceedings of the Human Language Technology Meeting (HLT-2002), San Diego, March 2002.

[4] S. Burger, V. MacLaren, H. Yu: The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style Proceedings of the ICSLP, Denver CO, September 2002.

[5] R. Stiefelhagen: Tracking Focus of Attention in Meetings. IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, October 14-16, 2002.

[6] D. Focken, R. Stiefelhagen: Towards Vision-based 3-D People Tracking in a Smart Room. IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, October 14-16, 2002.

[7] T. Schultz, A. Waibel, M. Bett, F. Metze, Y. Pan, K. Ries, T. Schaaf, H. Soltau, M. Westphal, H. Yu, and K. Zechner: The ISL Meeting Room System. Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001), Kyoto Japan, April 2001.

[8] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting browser: Tracking and summarizing meetings. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pages 281–286, Lansdowne, Virginia, February. 8-11 1998.

[9] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, A. Waibel, "Multimodal People ID for a Multimedia Meeting Browser," Proceedings of ACM Multimedia 99.

[10] X. Chen and J. Yang, Towards monitoring human activities using an omnidirectional camera, Proceedings of ICMI 2002, Pittsburgh, PA, October, 2002.

[11] R. Gross, J. Yang, A. Waibel, Face Recognition in a Meeting Room, Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG' 2000).