The “FAME” Interactive Space

Florian Metze, Petra Gieselmann, Hartwig Holzapfel, Thomas Kluge, Ivica Rogina, Alex Waibel, Matthias Wölfel, James L. Crowley, Patrick Reignier, Dominique Vaufreydaz, et al.

The “FAME” Interactive Space

F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, I. Rogina, A. Waibel, and M. Wölfel
Universität Karlsruhe (TH)
metze@ira.uka.de

J. Crowley, P. Reignier, and D. Vaufreydaz
Institut National Polytechnique de Grenoble (INPG)

F. Bérard, B. Cohen, J. Coutaz, and S. Rouillard
Université Joseph Fourier (UJF), Grenoble

V. Arranz, M. Bertrán, and H. Rodriguez
Universitat Politecnica de Catalunya (UPC), Barcelona

http://www.fame-project.org/

Abstract. This paper describes the “FAME” multi-modal demonstrator, which integrates multiple communication modes – vision, speech and object manipulation – by combining the physical and virtual worlds to provide support for multi-cultural or multi-lingual communication and problem solving.
The major challenges are automatic perception of human actions and understanding of dialogs between people from different cultural or linguistic backgrounds. The system acts as an information butler, which demonstrates context awareness using computer vision, speech and dialog modeling. The integrated computer-enhanced human-to-human communication has been publicly demonstrated at the FORUM2004 in Barcelona and at IST2004 in The Hague.
Specifically, the “Interactive Space” described features an “Augmented Table” for multi-cultural interaction, which allows several users at the same time to perform multi-modal, cross-lingual document retrieval of audio-visual documents previously recorded by an “Intelligent Cameraman” during a week-long seminar.

1 Introduction

Current advances in language and vision technology as well as user interface design are making possible new tools for human-human communication. Integration of speech, vision, dialog, and object manipulation offers the possibility of a new class of tools to aid communication between people from different cultures using different languages.

The “FAME” project (EU FP5 IST-2000-28323) develops a new vision for computer interfaces, which replaces and extends conventional human-computer interaction by computer-enhanced human-to-human (CEHH) interaction. The crucial difference lies in the role that the computer plays and the demands it makes on the human user’s attention in a living and working environment.

Like an invisible information butler, such systems observe and learn their users’ ways and preferences, and “understand” enough about the context and purpose of their activity, to be able to provide helpful and supportive services that are informed by, and appropriate for that context. A broad range of applications can profit from CEHH interfaces, including office work, communication, entertainment and many more. What is common in most of these settings is that the goal and preoccupation of visitors is to interact with other humans and not with machines.

The work presented here demonstrates an interactive space using intelligent agents to facilitate communication among researchers from different linguistic backgrounds, who discuss several scientific topics, which have been presented at a seminar series. The agents provide three services:

1. Provide information relevant to context, i.e. make users aware of existing information in an audio-visual database, then retrieve and present it appropriately
2. Facilitate human to human communication through multi-modal interaction as well as presentation of speech and text in multiple languages
3. Make possible the production and manipulation of information, blending both electronic and physical representations

The agents therefore do not explicitely intervene, but remain in the background to provide appropriate support (this mode is known as *implicit interaction*), unless explicitely called for to provide a particular service, such as playing a video.

The remainder of Section 1 will formulate the concepts behind "FAME" and introduce the functions our "intelligent room" [1] can perform. Section 2 will present the individual components and evaluate their performance individually, while Section 3 will present an overall evaluation of the integrated system from the user's point of view.

1.1 The Functions of the "FAME" Demonstrator

The functions of the "FAME – Facilitating Agents for Multi-Cultural Exchange" multi-modal demonstrator are split into an off-line and on-line part, as shown in Fig. 1:

Off-line Part. To provide audio-visual documents, an "Intelligent Cameraman" [2] recorded a four day seminar on Language Technology and Language, Cognition, and Evolution [3], which were held on the premises of the FORUM2004 [4] in Barcelona. The cameraman runs fully automatically and does not interfere with the lecture or its visitors.

The resulting videos, which contain audio from the presenter, the translator, and the audience, are then segmented, automatically transcribed and processed by a topic detection system, which assigns one topic to every segment. The videos can then be retrieved during the on-line use of the system in different ways.

On-line Part. The goal of the "Augmented Table" is to aid and support multi-cultural interaction and multi-modal document retrieval [5]. Students and researchers can come to the table and discuss with each other, retrieve information about the seminar as well as see recordings or automatically generated transcriptions of the lectures themselves, or see or give "testimonies".

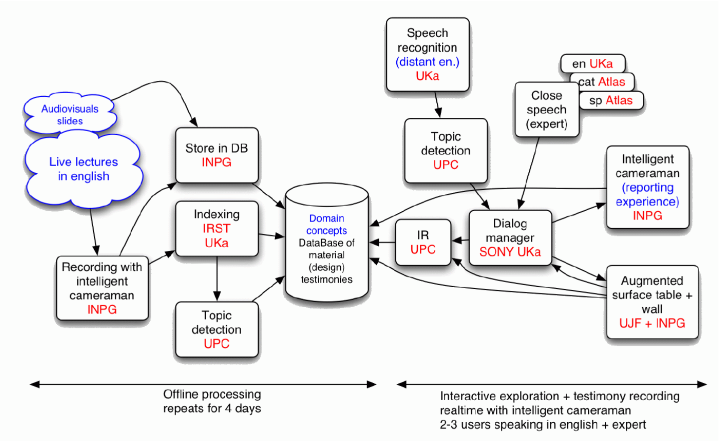
Speech recognition (distant en.) UKa
en UKa
cat Atlas
sp Atlas
Close speech (expert)
Audiovisuals slides
Store in DB INPG
Topic detection UPC
Live lectures in english
Intelligent cameraman (reporting experience) INPG
Indexing IRST UKa
Domain concepts DataBase of material (design) testimonies
IR UPC
Dialog manager SONY UKa
Recording with intelligent cameraman INPG
Augmented surface table + wall UJF + INPG
Topic detection UPC
Offline processing repeats for 4 days
Interactive exploration + testimony recording realtime with intelligent cameraman 2-3 users speaking in english + expert

Fig. 1. Components of the "FAME" demonstrator: the database connects the off-line "Intelligent Cameraman" and indexing to the left, the and on-line multi-modal document retrieval using the "Augmented Table" on the right

The Database. The database is the interface between the off-line and on-line parts. In addition to the videos, their transcription and topic segmentation, it contains speakers' CVs, a picture, and contact information. For every lecture, we also added information on time and place as well as course material (slides) and other meta information.

1.2 The "Interactive Space"

At a conference, attendees usually use the physical conference program organized as a time-schedule and they ask other attendees what's interesting to attend. Therefore, the design of our showcase must maintain these two forms of interaction (browsing a time-schedule and chatting with colleagues for reporting their experience and asking for advice) in a seamless and complementary way.

Preliminary Analysis. In a user-centered approach, domain concepts that users manipulate are classified to provide guidelines for the design of the user interface. In particular, *1st-class* concepts should be observable at all time, and *2nd-class* concepts should be browsable at all time, i.e. they are not necessarily observable but accessible through additional user's actions such as scrolling, selecting a navigation button such as "next-previous-top", "show details". After analysis of 15 CHI (Computer Human Interaction) conference programs, our demonstrator ranks 1st-class domain concepts as follows: *Lecture* (in terms of: Title, Place, Date, BeginHour, EndHour), *Speakers* and *Topics*; while we use the 2nd-class domain concepts *Documents* (papers, slides, video, audio) and *Testimonies* (recorded audio-visual "opinions" about lectures).

Design Solution. Given the current state-of-the-art in automatic speech recognition without close-talking microphones [6], our design is based on the hypothesis that there should be two types of users: one *manager* and up to five *users*:

Users are researchers knowledgeable in the task domain, i.e. they are skilled at manipulating the notions of session, lecture, topics; they are familiar with the lectures' topic, but not with augmented reality. To ensure a reliable functioning of the system, the users do not issue spoken commands to the system, but the topic spotter component of the system recognizes their speech and acts upon it. Visitors to the demonstration can be "users".

A manager is skilled both in the task domain and in the interaction techniques. His spoken requests are recognized by the system, so there is no need for extra explicit interaction with the system. He can conduct a rich dialog with the system using speech and direct manipulation, he acts as a "moderator".

The manager and several users can interact with the system at the same time, i.e. they can speak between themselves, move tokens or select items etc. As shown in Fig. 2, the overall setting includes:

- A shared horizontal surface (the "Augmented Table") for exploring information about the conference and recording testimonies
- A microphone in the middle for acquiring user's utterances and detecting information of interests (i.e. topics and/ or speakers' names) as well as a loud-speaker, so that the system can talk back to users
- A shared vertical surface (the "Wall") for peripheral activities, such as a presentation of the detected topics, lecture slides, or other text information from the database
- A second vertical surface for video projections from the database
- A camera to record testimonies

Compared to the "un-augmented" setting (i.e. browsing a paper brochure or asking colleagues in the hallway), the added value of FAME comes from (see Fig. 2):

- The interaction itself (manipulating tokens and chatting at the same time),
- The nature of the information that can be retrieved (lecture videos, presentation slides, testimonies, biographical or contact information, etc.),
- The capacity to produce feedback (e.g. reporting a testimony, which is added to the database), which improves the feeling of engagement in the event.

2 Components of the FAME Demonstrator

The components of the FAME demonstrator were developed in parallel by the FAME partners, each using their own software and tools. Integration of the individual components into one system was achieved during two integration workshops held at different sites using different hardware to provide backup and to ensure that the demonstrator would not depend on specifics of either location, so as to avoid problems during several days of public exhibits at the FORUM2004 in Barcelona and IST2004 in The Hague.

Communication between components is assured through Open Agent Architecture (OAA) [7], which allows agents running on several machines with different operating systems to communicate and also access the database. In total, the on-line part of the FAME demonstrator requires six PCs in a mixed configuration of Linux and Windows, which supported three speech recognizers, speech synthesis, video processing for the "Augmented Table", projection of the table itself and the walls, video player and recorder, room control, and database/ communication server.

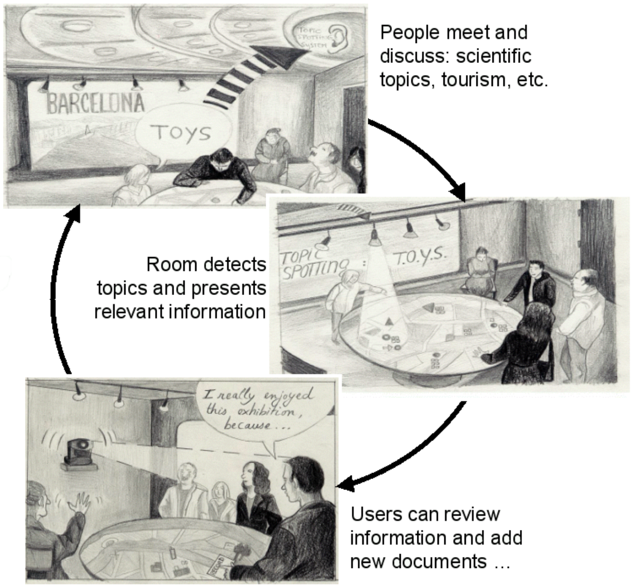
People meet and discuss: scientific topics, tourism, etc.
BARCELONA
TOYS
Room detects topics and presents relevant information
TOPIC SPOTTING : T.O.Y.S.
I really enjoyed this exhibition, because...
Users can review information and add new documents ...

Fig. 2. The overall setting of the "interactive space": table for focused search activity and walls for peripheral activities and information. The Feedback loop made possible by the FAME room improves engagement within communities

2.1 Cameraman and Speech Activity Detection

The automatic cameraman is a context aware system designed to record lectures or meetings. It creates movies by dynamically selecting camera views based on context analysis using a speech activity detection system running on several microphones.

The Context Aware Framework. This is based on the concept of entities, roles, relations, situations and a graph of situations [8]. An *entity* is a group of observable properties detected by perceptual components. For instance, a person (an entity with (x, y) properties) can be detected by a video tracker. A *role* is an acceptance test which selects entities based on their properties. For instance, a long and thin object can play the role of a pointer. A *relation* is a predicate over entities: person 1 is near person 2. A situation is defined as a set of roles and relations between entities.

Context is described as a graph. Nodes are situations, arcs are temporal constraints between the situations. They are decorated using Allen's temporal operators [9]. For each Allen temporal operator we composed a corresponding Petri Net pattern. By applying those patterns, the context graph can be transformed in an equivalent synchronized Petri Net. The Petri Net is then compiled in a set of forward production rules (Jess) that will check the situation activation sequence.

The Cameraman. The room is equipped with four cameras: a wide angle for a global view of the scene, one for the audience, one for the lecturer and one filming the slides. The lecturer has a lapel microphone and there are ambient microphones for the audience questions. The perceptual components are a video tracker [10] for the lecturer, a "new slide" detector based on image difference, and a speech activity detector. We model for situations, each with an associated camera: *lecturer speaks*, *audience asks a question*, *new slide*, and *unspecified*.

Speech Activity Detection (SAD). The SAD system is composed of several sub-systems: an *energy detector*, a *basic classifier* and a *neural net* trained to recognize voiced segments, e.g. vowels. The energy detector uses pseudo energy (sum of absolute sample values to determine if a signal is speech or not. The basic classifier works in frequency bands and tags specific sound classes: fricatives, low frequency sounds like computer or air conditioning fans, and other sounds. The neural net is a multi-layer perceptron with 2 hidden layers. Input consists of band crossing, energy and 16 LPC coefficients. This module is the only sub-system that needs to be trained. Training was done on 1 hour of speech extracted from the BREF [11] corpus. A rule based automaton determines the final result using the three sub-system's decision as input.

There are two evaluation conditions, one dealing with close-talking data from the lecturer's lapel microphone, the other dealing with far-field microphone audio. Evaluation results in terms of Mismatch Rate (MR), Speech Detection Error Rate (SDER), and Non-speech Detection Error Rate (NDER) are as follows [12]:

	MR	SDER	NDER
Close-Talking	10.8%	5.8%	27.9%
Far-Field	13.4%	12.9%	15.4%

These results and field experiments suggest that the SAD system accuracy is satisfying for the purpose of the context analysis.

2.2 Information Retrieval and Topic Detection

Information Retrieval (IR). Conventional IR allows a user to type a query, the system performs a search on a previously indexed single collection of information sources and then provides the results to the user. FAME needs to go beyond this scheme, because it not only deals with "conventional" IR, but also with on-line access to relevant documents from a *set of different collections* triggered by explicit or implicit interaction [13].

Indexing is done using textual features (words, stems, lemmas, multiword terms, phrases, etc.). These can also include morphological, syntactic or semantic information. Textual features can be automatically extracted from multimedia documents or manually attached to them.

Querying can be done using text or speech. The system allows cross-lingual and conceptual-based query expansion using EuroWordNet [14] as lexical resource. Cross-lingual IR is useful, because a speaker's ability to understand a document in a foreign language is usually higher than his active command of that language.

Topic Detection (TD). Detection of topics consists of assigning topics to lecture segments, or detecting that no in-domain topic is currently being discussed.

In order to perform (on-line) TD, topics first need to be described during a preparation phase. We refer to this off-line and highly time-consuming task as Topic Characterization (TC). The most widely used form of performing TC consists in attaching a collection of units to each topic. The problems which need to be addressed here are *selecting* the appropriate type of unit, *choosing the criteria* for selecting the appropriate units to describe each topic, and *choosing a representation schema* for the collection of units describing the set of topics.

Topic Signatures (TS) are used to represent topics. TS are lists of weighted units or terms. Building a TS for a collection of topics was done semi-automatically. Different TS for a topic collection depend on each other.

With the set of ”topics” pre-defined and each topic being described by its TS, the input stream is examined sequentially and its features are extracted and compared with the features (TS) of the different topics. If there is enough evidence, the presence of a new topic (a change of topic) is detected and communicated [15].

2.3 The Augmented Table and Token Tracker

The token tracker component of the “Augmented Table” analyzes in real time the video stream coming from the camera located above the table. The positions of all tokens (red disks) are extracted from the video stream and converted into a stream of APPEAR, MOTION and DISAPPEAR events with associated positions (x, y). The FAME token tracker will be evaluated in terms of its *resolution* and *latency*.

Resolution. The smallest motion that will be detected and reported by the tracker depends on the video stream, which is processed at 388 x 284 pixels in order to maintain full video frame rate and low latency. Setting a 2 pixel motion threshold for static stability before reporting a motion event because of the instability of the video signal, the actual resolution on the surface depends on the geometry of the setup. Typically, an accuracy of 0.5 cm or 5 pixels for the projection, is reached, which is sufficient for the task.

Latency. The time lag between the moment a user moves a token and the moment the associated feedback is moved accordingly has a strong effect on human performance with interactive systems. [16] recommends that devices should not have more than 50 ms latency. As shown in [17], it is not necessary for latency to be very stable: no effect on human performance is detected with standard deviation of latency at or below 82 ms.

Following the approach reported in [18], we designed a latency measuring device for the token tracker: on a 1.4 Ghz PowerPC G4 machine, the resulting latency value is distributed between 61 ms and 141 ms, with an average at 81 ms, which is tolerable given a standard variation on latency of 16 ms in our system, which means FAME is well within the 82 ms requirement expressed in [17]. We mainly attribute the latency variation to the difference in camera delay (25 Hz frame rate) and projector refresh rate (60 Hz).

2.4 Dialog Manager

The dialog manager is the central on-line component that mediates and executes messages of the various components, maintaining a shared multi-modal context with the “Augmented Table”. In this function, the dialog manager interprets speech input from the manager (*explicit interaction*), input from the topic spotter using distant speech recognition (*implicit interaction*) and events stemming from users’ actions on the table (*explicit interaction*). The dialog manager will call requested services as dictated by the context, for example IR queries, highlighting information on the table, playing video segments, showing slides or pieces of information on the information wall, controlling testimony recording, controlling room devices (lamps) and generating spoken output.

If the dialog manager detects clarification needs, or information is missing to execute a command, it requests further information from the speaker or calls on other services, such as IR, to obtain the information required to execute the command.

For dialog management and multi-modal integration we use the TAPAS dialog management framework, which is based on ARIADNE dialog algorithms [19] with extensions to support multilingual input and output [20]. The dialog manager can restrict the search space of the close-talking speech recognizers to the given context. Performance numbers for these recognizers and the dialog manager are summarized here:

Performance on	Spanish	English
Overall Number of Turns	597	1029
Word Error Rate	18.6%	17.2%
Sentence Error Rate	21.3%	20.7%
Turn Error Rate	20.7%	17.5%
Finalized Goal Rate	75.2%	71.5%

2.5 Speech Recognition

Apart from the close-talking recognizers for the managers, the augmented table uses two other speech recognition systems:

Lecture Transcription. To automatically segment and transcribe the lectures given at the seminar and make them accessible to indexing, transcriptions were generated by ITC-irst [21]; here we describe a smaller backup system developed by UKA, which was run during pauses between demonstrations. Development of lecture transcription components used the TED [22] corpus, because of its similarity to the Barcelona seminar presentations. The acoustic model has been trained on 180 h Broadcast News (BN) and close-talking “Meeting” data [6], summing up to a total of 300 h of training material. Models were first trained on Meeting and BN, then adapted to TED by MAP and MLLR.

To generate language models (LMs), we used corpora consisting of BN (160 M words), proceedings (17 M words) of conferences such as ICSLP, Eurospeech, ICASSP or ASRU and talks (60 k words) by the TED adaptation speakers. The baseline LM is based on BN. Our final LM was generated by interpolating a 3-gram LM based on BN and proceedings, a class based 5-gram LM based on BN and proceedings and a 3-gram

LM based on the talks. The overall out-of-vocabulary rate is 0.3% on a 25 k vocabulary including multi-words and pronunciation variants.

The TED decoding and perplexity results using the acoustic and language models described above are shown below. The first run did not use TED model adaptation, while the last run used MLLR model adaptation on the hypothesis of the "TED-Adapted" run.

Performance on	Native		Non-native	
TED database	WER	PP	WER	PP
First run	42.8%	300	53.9%	160
TED Adapted, Baseline LM	36.7%	300	35.2%	160
Adapted to Speaker, Final LM	28.5%	171	31.0%	142

Comparing the first runs between the native vs. the non-native test data sets we see a big gap in performance which is due to the fact that the acoustic model was trained on mostly native speech. The gain of acoustic adaptation was higher for the non-native test set as for the native test set, which may be explained by the fact that the amount of adaptation material was four times smaller for the native speakers than for the non-native speakers. Word error rates of transcriptions of the Barcelona lectures are about the same as those of the TED development set.

Topic Spotting using Room Microphones. A real-time speech recognition system derived from the ISL "Meeting" system [6], running on a single microphone placed in the middle of the "Augmented Table" is used to detect topics from the conversation between visitors to FAME.

The initial system using a BN LM interpolated with data collected from the Internet had a topic error rate of 41% at a speech recognition accuracy of 17%. Re-weighting the keywords for topic spotting and taking confidence scores into account reduced the topic spotting error rate to 38%.

2.6 Component Integration

We have conducted an evaluation to measure response time over a distributed OAA system including different components. OAA offers two kinds of method calls: *asynchronous* calls (messages without response) and *synchronous* calls (response is returned to the caller). We have evaluated both message types with sequential calls and parallel calls for maximum throughput.

We found that the first call to OAA of an agent plus registration time was quite large with an average of 667 ms. All following calls are much faster. The average delay in our distributed system for a single synchronous method call (including request and reply) ranges between 20 ms and 25 ms. The numbers have been evaluated for an agent calling himself (routed over the OAA facilitator), chained calls (a call stack of depth 10), and a single agent calling another agent. The same setting has been used to create parallel tests, where 10 agents send a message concurrently. The average response time was 157 ms. The delay from sending the first request and receiving the last response was 207 ms. Divided by 10, this corresponds to an average processing time of 21 ms per message. We thus assume the total message throughput to be 47 messages (including responses) per second. The results of the sequential tests and sequential chained calls correspond to the final setting of the demonstrator and provide realistic numbers of the expected delay of messages. In the demonstration setup, the delays of the messages are small enough. However, the tested setup can not be applied to a larger system or one that needs a significantly higher message throughput.

3 The FAME Demonstrator and User Study

Over 150 persons used FAME during the Barcelona demonstration. We randomly selected 5 groups of 3 users among them for a user study [23]. After a brief introduction to the system, users were asked to perform predefined tasks, such as "Can you retrieve Mr. Harold Somer's lecture?" or "What topics are being addressed in the lecture about Machine Translation?", etc. We observed their behavior while interacting within the FAME interactive space and finally, the users answered some questions about their impression of the system.

We observed two general kinds of behavior: *actors* who wanted to experiment with the system, and *spectators* who enjoyed observing the scene. People felt confident with the system in less then one minute. They were more interested in playing with it than learning from the content: unfortunately, the social context did not favor a "learning" experience, although most visitors had a "science" background. We summarize the results in two categories:

General Impression. The function and the manipulation of tokens are easy to learn and to understand. The existence of multiple tokens brings a playful note to the system. We observed users dropping multiple tokens randomly on the table to see what happens. The multiplicity of tokens supports simultaneous actions but does not necessarily favor collaboration. In turn, this lack of collaboration between users may lead to confusion, because, as a result of the quasi-simultaneity of different output modalities, users did not know whether the replies of the system corresponded to their own or someone else's request.

Most users would have preferred to be able to ask IR requests directly, instead of asking the manager to do it for them. Most users did not pay attention to the fountain of words (on the wall) that correspond to the topics recognized by the topic spotter as users were talking freely around the table. From the social perspective, the testimony function was a big success. At first, some users were shy at recording themselves. Others enjoyed it for the possibility to watch themselves through the replay function.

General Impression	Yes	Sometimes	No	No Answer
Is it useful to have multiple places to look for information?	5	8	1	1
Is it fun to have multiple places to look for information?	7	6	0	2
Is it useful to be able to play with the system with other people?	9	5	0	1
Is it fun to be able to play with the system with other people?	11	3	0	1
Would you prefer to issue speech commands yourself?	7	4	2	2
Is the system reliable?	12	0	2	1
Would you be interested in using it?	12	1	0	2

User Ratings For Different Parts of the System. In general, users' opinions are quite enthusiastic, but overall, the system is viewed as a prototype. Some users judged the quantity of information insufficient to really test the system. Some videos were perceived as too short but the content was supposed to be relevant. The icons and numbers denoting video clips in the flower menus should be replaced by a representative image of the video content and a textual title. Multi-surface and multi-user interaction were considered useful and fun but sometimes confusing. In addition, users would like to navigate within a video clip. Some users complained about the speed of the system reactions.

Rating on	Very Easy	Easy	Difficult	Very Hard	No Answer
Working with the tokens	1	13	0	0	1
Understanding organization of the seminar, topics and speakers	0	13	1	0	1
Choosing lectures to visit, discovering topics and speakers	2	11	1	0	1
Retrieving audio-visual content from lectures	1	13	0	0	1
Retrieving course materials	2	11	1	0	1
Giving a testimony	3	10	0	1	1

In summary, the overall design and technical integration of the FAME interactive space were very well perceived, fun to use, and provided a special experience. However, at the detail level, there is room for both design and technical improvements, such as speeding up the response time, allowing easier navigation within videos and slides, and better spatial arrangements for projection walls and augmented table, etc.

4 Conclusions

From the interaction perspective, the system is easy to learn and provides people with an enjoyable experience based on the tangibility of the tokens, the table, the walls and spoken interaction. The novelty of the interaction style and the fact that it is fun, draw users' attention away from the content itself. The “testimony” feedback service provides a new social experience, although, in our experiment, users exploited this facility more to look at themselves and not to produce information for others.

We have not been able to demonstrate the benefit of implicit interaction supported by the topic spotter. We believe that implicit interaction cannot be tested in sessions of short duration, but can only be appreciated over a long period of use, which was not possible in our setting.

The experiment also showed the necessity for systems to automatically adapt to the physical environment: spatial relationships between surfaces, distance from the actuators, background color of surfaces, etc. For example, one third of the users were wearing the same color as that of the tokens, which could be a problem in bad lightning conditions. Our continuing work on context modeling and adaptation is an attempt to solve this problem.

References

1. Gieselmann, P., Denecke, M.: Towards multimodal interaction within an intelligent room. In: Proc. Eurospeech 2003, Geneva; Switzerland, ISCA (2003)
2. Crowley, J.L., Reignier, P.: Dynamic composition of process federations for context aware perception of human activity. In: Proc. International Conference on Integration of Knowledge Intensive Multi-Agent Systems, KIMAS'03, 10, IEEE (2003)
3. Consorci Universitat Internacional Menéndez Pelayo de Barcelona: ”Tecnologies de la llengua: darrers avenços” and ”Llenguatge, cognició i evolució”. http://www.cuimpb.es/ (2004)
4. FORUM2004: Universal Forum of Cultures. http://www.barcelona2004.org/ (2004)
5. Lachenal, C., Coutaz, J.: A reference framework for multi-surface interaction. In: Proc. HCI International 2003, Crete; Greece, Crete University Press (2003)
6. Metze, F., Fügen, C., Pan, Y., Waibel, A.: Automatically Transcribing Meetings Using Distant Microphones. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Philadelphia, PA; USA, IEEE (2005)
7. SRI AI Center: Open Agent Architecture 2.3.0. http://www.ai.sri.com/˜oaa/ (2003)
8. Rey, G., Crowley, J.L., Coutaz, J., Reignier, P.: Perceptual components for context aware computing. In: Proc. UBICOMP 2002 – International Conference on Ubiquitous Computing, Springer (2002)
9. Allen, J.: Towards a general theory of action and time. Artificial Intelligence **13** (1984)
10. Caporossi, A., Hall, D., Reignier, P., Crowley, J.: Robust visual tracking from dynamic control of processing. In: PETS04, Workshop on Performance Evaluation for tracking and Surveillance, ECCV04, Prague; Czech Republic (2004)
11. Lamel, L., Gauvain, J., Eskenazi, M.: BREF, a large vocabulary spoken corpus for French. In: Proc. Eurospeech 1991, Geneva, Switzerland (1991)
12. Surcin, S., Stiefelhagen, R., McDonough, J.: Evaluation packages for the first chil evaluation campaign. CHIL Deliverable D4.2 (2005) http://chil.server.de/.
13. Bertran, M., Gatius, M., Rodriguez, H.: FameIr, multimedia information retrieval shell. In: Proceedings of JOTRI 2003, Madrid; Spain, Universidad Carlos III (2003)
14. The Global WordNet Association: EuroWordNet. http://www.globalwordnet.org/ (1999)
15. Arranz, V., Bertran, M., Rodriguez, H.: Which is the current topic? what is relevant to it? a topic detection retrieval and presentation system. FAME Deliverable D7.2 (2003)
16. Ware, C., Balakrishnan, R.: Reaching for objects in vr displays: Lag and frame rate. ACM Transactions on Computer-Human Interaction (TOCHI) **1** (1994) 331–356
17. Watson, B., Walker, N., Ribarsky, W., Spaulding, V.: The effects of variation of system responsiveness on user performance in virtual environments. Human Factors, Special Section on Virtual Environments **3** (1998) 403–414
18. Liang, J., Shaw, C., Green, M.: On temporal-spatial realism in the virtual reality environment. In: ACM symposium on User interface software and technology, Hilton Head, South Carolina (1991) 19–25
19. Denecke, M.: Rapid prototyping for spoken dialogue systems. In: Proceedings of the 19th International Conference on Computational Linguistics, Taiwan (2002)
20. Holzapfel, H.: Towards development of multilingual spoken dialogue systems. In: Proceedings of the 2nd Language and Technology Conference. (2005)
21. Cettolo, M., Brugnara, F., Federico, M.: Advances in the automatic transcription of lectures. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), Montreal; Canada, IEEE (2004)
22. Lamel, L., Schiel, F., Fourcin, A., Mariani, J., Tillmann, H.: The translanguage english database (ted). In: Proc. ICSLP1994, Yokohama; Japan, ISCA (1994) 1795 – 1798
23. Coutaz, J., et al.: Evaluation of the fame interaction techniques and lessons learned. FAME Deliverable D8.2 (2005)