IDIAP COMMUNICATION

The IDIAP Smart Meeting Room

Darren C. Moore [a]

IDIAP–Com 02-07

November 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] darren.moore@idiap.ch

1 Abstract

The IDIAP Smart Meeting Room is a meeting room equipped with synchronised, multi-channel audio-visual recording facilities. This document presents a detailed description of the room with particular emphasis on the acquisition equipment and the components used to synchronise and accurately time-stamp each channel of audio and video. Brief descriptions of the current meeting recording configuration and recorded data processing procedures are also included.

2 Introduction

The IDIAP Smart Meeting Room was installed for the purpose of acquiring audio, visual and textual data within meeting scenarios. These recordings are needed to support the wide range of speech, image and multi-modal research efforts that occur at IDIAP, and at other partner institutions.

Section 3 of this document presents an overview of the meeting room configuration and capabilities. Sections 4, 5 and 6 describe the technical details of the audio and video acquisition equipment and the components used to achieve accurate syncronisation between multiple channels of audio and video. Section 7 describes the interconnections between components and how they have been installed in the Smart Meeting Room. A brief discussion of how meetings are currently recorded, and the post-processing that is performed on the acquired data is presented in Sections 8 and 9. Section 10 describes enhancements to the meeting room hardware that are being considered for future implementation.

3 Overview

The IDIAP Smart Meeting Room is a 8.2m×3.6m×2.4m rectangular room containing a centrally located 4.8m×1.2m rectangular table (suitable for seating 12 people). A white-board and retractable projector screen occupy the wall at one end of the room, while the other end of the room houses a 19" rack with the audio-visual acquisition equipment, a small table for a telephone, and an air conditioning vent. Metal rails have been installed on all outer edges of the ceiling to allow flexible placement of cameras and lights. A beamer has been suspended from the ceiling above the meeting table.

The Smart Meeting Room has 24 miniature lapel microphones, which can each be used either as a tethered lapel microphone attached to a meeting participant, or as part of tabletop microphone arrays. All microphone inputs can be acquired simultaneously. The Smart Meting Room currently has three video cameras that can be suspended from the ceiling, or placed on tripods at any location around the room. All three video channels can also be simultaneously recorded.

All cabling for the microphones and cameras has been professionally installed, and runs out-of-sight in channelling beneath the table and around the walls. Each seating position around the meeting table has a microphone socket, making the use of tethered lapel microphones easier and more comfortable for meeting participants.

4 Audio Acquisition

The audio acquisition equipment consists of the following,

- 24 Sennheiser MKE 2-5-C miniature electret microphone
- 1 Custom-built microphone power box
- 3 PreSonus Digimax preamplifier/digitiser
- 1 Mark of the Unicorn 2408mkII PC interface
- 1 Win2k PC with Cakewalk SONAR recording software

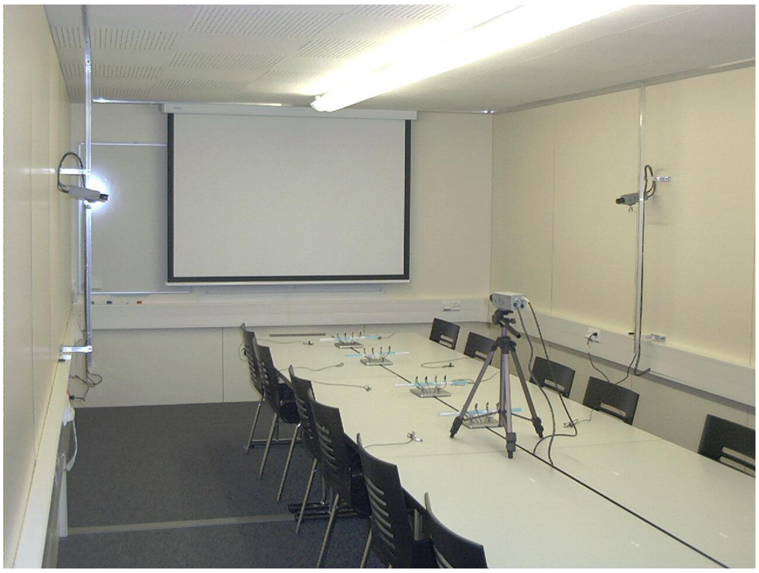

Figure 1: IDIAP Smart Meeting Room

- 1 Sennheiser EW112 wireless microphone setup

Each component is described in detail below.

4.1 Sennheiser MKE 2-5-C

The Sennheiser MKE 2-5-C [1] is a high quality miniature electret microphone. It has an approximately linear frequency response between 20Hz and 20kHz and omni-directional characteristics. It has a very high sensitivity of 31mV/Pa, making it a good choice for use in microphone array applications where the desired source is not in the immediate vicinity of the array. The diameter of the microphone is 6mm.

4.2 Custom Microphone Power Box

The disadvantage of the MKE 2-5-C is that requires a separate DC bias voltage in order to operate. It cannot be plugged directly into the microphone preamplifiers.

Due to the large number of microphones and the high cost of biasing modules produced by Sennheiser, it was decided to build a custom power box into which all microphones were input. A 12V DC adaptor is used with appropriate circuitry to supply each microphone with the correct biasing voltage. The power box also merges the cabling for all microphones into a 12m multicore cable that runs to the rest of the audio acquisition equipment. The use of the multicore cable considerably simplifies the microphone cabling requirements and increases the freedom of microphone placement.

4.3 PreSonus Digimax

The PreSonus Digimax [2] is a high-quality, 8-channel microphone preamplifier, with integrated 24-bit digitisation. Separate gain controls are provided for each input channel and the digital output is via a single ADAT lightpipe (optical fibre) containing all 8 channels. The Digimax is able to sample at rates of 32kHz, 44.1kHz and 48kHz, using either its own internal clock, or a wordclock supplied by an external source. The Digimax uses an external power supply to ensure noise-free amplification and digitisation across all input channels.

Figure 2: Sennheiser MKE 2-5-C (with tie-clip)

Figure 3: Custom microphone power box (front and rear)

Figure 4: PreSonus Digimax

Figure 5: Mark of the Unicorn 2408mkII

The 3 Digimax units in the Smart Meeting Room are housed in a 19” rack. The 24 input channels in the multicore cable from the microphone power box are fanned out inside the rack and connected to the individual inputs of the 3 Digimax units.

4.4 Mark of the Unicorn (MOTU) 2408mkII

The MOTU 2408mkII [3] is a flexible interface for PC hard disk-based audio recording. It consists of a 19” rack-mounted I/O unit connected via a firewire-like interface to a PCI card installed in a PC. The I/O unit has 24 inputs and outputs, divided into three banks of 8-channels. Each bank has a range of input and output options, including analog, S/PDIF, TDIF, or ADAT lightpipe and all inputs and outputs can operate simultaneously. The PCI card can accommodate up to three I/O units, allowing up to 72 simultaneous input and output audio channels. Driver software installed on the PC containing the PCI card allows configuration and acquisition to be controlled through software.

In the Smart Meeting Room acquisition system, the ADAT lightpipe outputs from the 3 PreSonus Digimax units are connected to the 3 input banks on a single MOTU 2408mkII I/O unit.

4.5 Win2k PC

The PC used with the MOTU 2408mkII needed to be reasonably high performance in order to support the continuous recording of 24-channels at sampling rates of 48kHz per channel for extended periods of time. The operating system needed to be a Windows variant for compatibility with the MOTU 2408mkII driver software, and to provide increased options for audio recording applications. The PC used in the Smart Meeting Room has the following specifications:

- Intel Pentium IV 2GHz CPU
- 512Mb RAM
- 2×80Gb IDE hard disc drives
- Windows 2000 Professional
- DVD writer
- IEEE1394 PCI interface card
- 19” rack-mount case

Figure 6: Sennheiser EW112 wireless receiver and transmitter (incl. microphone)

The DVD writer allows recordings to be easily transferred to DVD for archival or distribution to other project partners. The IEEE1394 interface card allows video recordings to be transferred from the Sony video cassette recorders (see Section 5.2).

The recording software used to control audio recording is Cakewalk SONAR [4]. A graphical user interface allows the user to easily configure recording sessions on the MOTU 2408mkII and to view and export the acquired data. It is also compatible with the synchronisation hardware described in Section 6.

4.6 Sennheiser EW112 Wireless Microphone

The tethered lapel microphones used for the meeting participants restrict them to remain relatively stationary for the duration of a meeting. If part of a meeting involves a participant giving a presentation or writing on the whiteboard, then the tethered microphones are not suitable.

For this reason, a single Sennheiser EW112 wireless microphone [5] was added to the Smart Meeting Room. The wireless microphone consists of a Sennheiser electret microphone (similar to those described in Section 4.1) connected to a small UHF transmitter that is worn on the body of a meeting participant. A receiver located on the 19” rack demodulates the UHF signal from the transmitter into the original baseband audio signal. If the wireless microphone is used, the receiver output is connected to a PreSonus Digimax input channel, overriding the input from one of the tethered lapel microphones.

5 Video Acquisition

The video acquisition equipment consists of the following:

- 3 Sony SSC-DC58AP CCTV cameras
- 3 Sony GV-D1000E digital video cassette recorder

Each component is described in detail below.

5.1 Sony SSC-DC58AP

The Sony SSC-DC58AP [6] is a high quality closed-circuit television (CCTV) camera. It has an internal power supply allowing it to be plugged directly into a 220V power outlet. A number of image

Figure 7: Sony SSC-DC58AP

adjustment features including Automatic White Balance (AWB) and Automatic Gain Control (AGC) are provided, allowing the camera to automatically adapt to different lighting conditions.

The SSC-DC58AP outputs both Y/C and composite video signals via S-Video and BNC output connectors respectively, and accepts a video synchronisation signal (eg. blackburst) through a second BNC connector.

In the Smart Meeting Room, each camera has been fitted with a wide angle lens that has a field of view of 38° to 80° and manual zoom, focus, and aperture controls. This allows greater flexibility in the placement of cameras around the room due to the relatively narrow room width.

5.2 Sony GV-D1000E

The Sony GV-D1000E video walkman [7] is a portable MiniDV video tape recorder with an in-built 4" LCD screen. It has s-video input and output sockets allowing it to be easily connected to the CCTV cameras for recording, or to the beamer for playback. An IEEE1394 interface is also provided for transferring video to a PC.

The Sony GV-D1000E was selected instead of a PC-based video capture solution because video recordings for all cameras needed to be full PAL resolution (720×576) and frame-rate (25fps). A single PC would only be able to support recording of one video channel at full PAL specifications due to the high data rates involved. Therefore, a PC solution to record the output of the 3 CCTV cameras would require three separate PC's. This would have been inefficient, unwieldy, and significantly more expensive.

The LCD screen on each video walkman has proved to be invaluable for calibration of cameras (position, focus, zoom) and for online monitoring of each camera view during meeting recording.

The disadvantage of this video capturing technique is the additional (manual) work required to transfer the video recordings from MiniDV tape to the hard drive of a PC for use in image processing research. However, MiniDV tapes do offer an efficient method of backup/archival of recordings.

6 Synchronisation

If the audio and video acquisition equipment described in Sections 4 and 5 was used without any further thought about synchronisation, the following inaccuracies would be present in the system,

- The Presonus Digimax units would take audio samples at different points in time, using internal clocks that varied slightly across units. The 24 audio channels would therefore not be synchronised with respect to each other.
- The Sony CCTV cameras would acquire frames of video at different points in time. In the worst case there would be ±20ms time difference between video channels.

Figure 8: Sony GV-D1000E

- Highly accurate re-merging of recorded audio and video channels would be impossible due to the complete lack of global timestamp information in all recorded streams, coupled with the inability to start all video and audio recording at *exactly* the same instant.

The following items of equipment were added to the Smart Meeting Room to alleviate these inaccuracies,

- 1 Horita BSG-50 PAL blackburst generator
- 1 Mark of the Unicorn MIDI Timepiece AV
- 3 Horita AVG-50 LTC-VITC translator

6.1 Horita BSG-50 PAL

The Horita BSG-50 PAL blackburst generator [8] produces a composite video reference signal and outputs this signal on multiple outputs. Other items of video and synchronisation equipment are then able to lock to this reference signal, therefore ensuring perfect synchronisation across all system components.

In the Smart Meeting Room, the BSG-50 acts as a master clock source, providing a common blackburst reference signal to each of the Sony SSC-DC58AP cameras. The cameras are able to synchronise with the blackburst signal, resulting in all cameras sampling video frames at *exactly* the same instant. Another BSG-50 output is connected to the MOTU MIDI Timepiece AV (described below), which locks to the blackburst reference signal and generates all other timing signals. Four blackburst outputs are unused, and are available for future expansion of the system.

6.2 MOTU MIDI Timepiece AV

The MOTU MIDI Timepiece AV [9] is a flexible synchroniser capable of slaving to and/or generating a variety of timing signals.

Figure 9: Mark of the Unicorn MIDI Timepiece AV

In the Smart Meeting Room, the blackburst output from the Horita BSG-50 PAL is input to the MIDI Timepiece AV, which locks to the reference signal and then uses it to derive the following timing signals,

- Word clock. This is a square wave signal that is used to synchronise audio sampling across different equipment.
- Longitudinal Time Code (LTC). This is an industry standard hours-minutes-seconds-frames (HH:MM:SS:FF) timecode where the frame count (FF) corresponds to a video standard (eg. PAL=25fps). For each video frame, the complete timecode is encoded into an 80-bit word, resulting in an 2kHz signal that can be recorded and played-back like any other audio signal.
- MIDI Time Code (MTC). This is a timing signal used by MIDI equipment, and contains the same time code as for LTC.

In the Smart Meeting Room, the MIDI Timepiece AV generates a 48kHz Word Clock which is transmitted to the external clock input of each PreSonus Digimax. This ensures that all 24 input audio channels are sampled simultaneously.

The LTC generated by the MIDI Timepiece AV is converted internally to MIDI Time Code (MTC) and transmitted via USB to the PC. The Cakewalk SONAR recording software on the PC is able to synchronise with the MTC and then accurately timestamp the acquired audio samples.

The LTC is also output directly to each Horita AVG-50 LTC-VITC translator for insertion into each video camera signal. This is discussed in the following section.

6.3 Horita AVG-50

The Horita AVG-50 LTC-VITC Translator [10] inserts a Vertical Interval Time Code (VITC) into an input video signal. The VITC can be internally generated, or can be derived from an incoming LTC signal. VITC is a television industry time code standard that encodes the same HH:MM:SS:FF information used in LTC into a 90-bit word. The 90-bit pattern is then inserted into the top few lines of each frame of video as a sequence of 90 black and white dots.

Normally, the VITC is inserted in the vertical blanking region of a video frame (ie. not visible on normal television screens). However, because MiniDV video recorders remove the vertical blanking region from a video signal prior to digitisation, the AVG-50 can be configured to insert the VITC into the active region of each frame. The AVG-50 can also optionally insert a character version of the VITC at the bottom of each frame, which is useful when manually viewing and editing recorded video.

In the Smart Meeting Room, the s-video output of each CCTV camera is input to separate AVG-50's along with the LTC generated by the MOTU MIDI Timepiece AV. Each AVG-50 converts the LTC to VITC and inserts it into the video frames. The s-video outputs from the AVG-50's (containing the VITC) are then fed to the s-video inputs on the Sony GV-D1000E MiniDV video recorders.

The VITC is guaranteed to be perfectly synchronised with each frame of video, because the cameras and the LTC source (ie. the MIDI Timepiece AV) are genlocked to the common video reference signal output from the Horita BSG-50 PAL.

The result of the VITC insertion is that each frame of video contains a time-stamp that corresponds exactly to the time-stamp that is used by Cakewalk SONAR and stored with the acquired audio. This

Figure 10: Horita AVG-50 (×3)

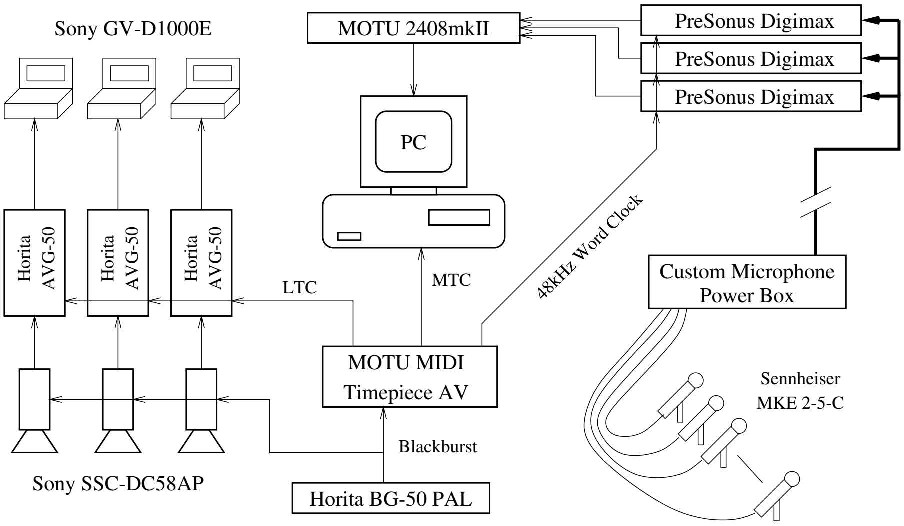

Figure 11: Smart Meeting Room recording equipment and connectivity

time-stamp can be easily interpreted in order to accurately realign any audio channel with any video channel.

7 Equipment Summary

The complete audio and video recording system including inter-component connectivity is illustrated in Figure 11. In the Smart Meeting Room, all components except the cameras, microphones and microphone power box are contained in a 19" rack located at the rear of the meeting room. A photograph of the rack is included in Figure 12.

8 Meeting Recording

The Smart Meeting Room is currently configured for full audio-visual recording of meetings with up to 6 participants. Two cameras have been suspended from the ceiling on either side of the meeting table, each providing a medium angle front-on view of three meeting participants. The third camera has been mounted on a tripod that is located on the far-end of the meeting room table, providing a view of the entire meeting including the whiteboard and projector screen.

The 24 microphone channels have been split between table-top microphone arrays and lapel microphones for the participants. Two circular microphone arrays each containing eight microphones are centrally located on the table between groups of 4 participants.

Figure 13 shows a top-down representation of the audio-visual meeting recording configuration.

For audio-only meeting recordings, the Smart Meeting Room can accommodate up to 12 meeting participants with lapel microphones, with the remaining microphones available for use in tabletop

Figure 12: Meeting room recording rack

microphone arrays.

9 Data Processing

The output of a meeting recording session using the configuration and equipment in Figures 13 and 11 is 24 mono wave files on the hard drive of the meeting room PC and 3 MiniDV cassettes. The video on each MiniDV cassette is manually transferred to AVI files on the meeting room PC via the IEEE1394 interface. The DV compression used on the MiniDV cassettes is maintained in the acquired AVI files. This results in approximately 12Gb of AVI-format video per camera per hour. Each audio wave file is downsampled from 48kHz to 16kHz, resulting in approximately 110Mb of audio data per microphone per hour. Thus the total amount of audio-visual data generated by a one hour recording is approximately 38.6Gb, which is equivalent to a data rate of 11Mb per second.

The initial meeting recordings in the Smart Meeting Room have been archived onto DVD-R. However, 9 DVD-R's are required to archive one hour of meeting recordings, which quickly becomes unwieldy, especially for the distribution of data between project partners. To alleviate these problems, IDIAP and other members of the IM2 research network are currently implementing a high capacity "media fileserver" to archive audio-visual recordings and provide project partners with internet access to the data. A first phase implementation is expected to be completed in January 2003.

10 Future Enhancements

The audio, video and synchronisation equipment is entirely scalable, making the future integration of additional microphones and cameras a straightforward task.

A binaural manikin (see Figure 14) is on loan from the University of Sheffield, and will soon be integrated into the audio recording system. The manikin will utilise two audio channels that are currently unused in the 6-person audio-visual recording configuration (see Section 8).

It is desirable to record the output of the beamer, so that presentation material (slides, movies,

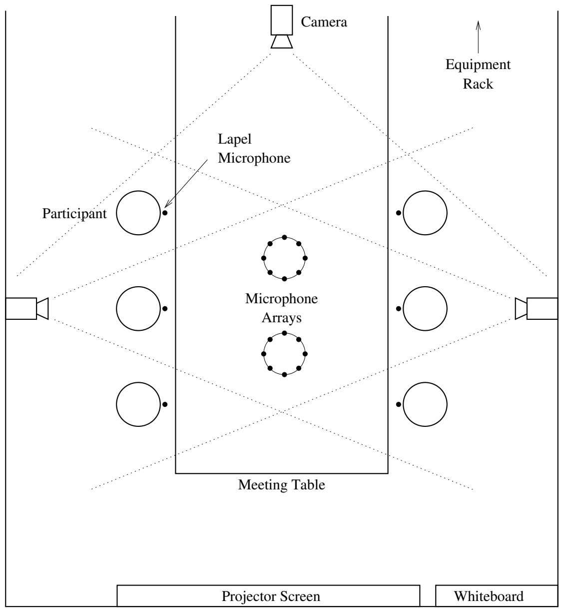

Figure 13: Smart Meeting Room audio-visual recording configuration

Figure 14: Binaural manikin

demonstration applications, etc) can be archived with the audio and video recordings. The simplest solution is to convert the RGB output of the beamer into an s-video signal that gets treated in exactly the same manner as the outputs from the CCTV cameras (ie. insert VITC, record on MiniDV cassette). This approach accurately synchronises the beamer output with all other audio and video channels, but is horrendously inefficient in terms of the storage required. The other downside is that the extraction of text from presentation slides becomes a difficult OCR problem. One alternative is to use a framegrabber instead of a digital tape recorder, and acquire frames at less frequent intervals.

Other possibilities for future enhancement of the Smart Meeting Room capabilities include:

- close-up cameras for all meeting participants
- document camera
- integrated teleconferencing/videoconferencing facilities
- pan-tilt-zoom cameras
- cameras dedicated to capturing the meeting table surface

11 Conclusion

The IDIAP Smart Meeting Room is capable of recording high quality, multi-channel, audio-visual meeting data. The current configuration uses three cameras with wide-angle lenses, six lapel microphones, and two 8-element microphone arrays to record meetings containing up to six participants. All channels of audio and video are accurately synchronised and time-stamped. The system has been designed with scalability in mind to permit straightforward expansion of the video and audio capabilities in the future.

12 Acknowledgements

Many staff at IDIAP have been involved in the installation and operation of the Smart Meeting Room. These people are: Thierry Collado, Olivier Masson, Iain McCowan, Daniel Gatica-Perez, Pierre Wellner and Frank Formaz.

The author wishes to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)".

This work was also funded by the European project "M4: MultiModal Meeting Manager", through the Swiss Federal Office for Education and Science (OFES).

References

[1] Sennheiser Electronics. *MKE 2 Instructions For Use*, 1995.

[2] PreSonus Corporation. *Digimax User's Manual.*

[3] Mark of the Unicorn Inc. *MOTU 2408mkII User's Guide.*

[4] Cakewalk. *SONAR Digital Multitrack Recording System User's Guide*, 2001.

[5] Sennheiser Electronics. *Evolution Wireless Series EW100 Instruction Manual*, 2001.

[6] Sony Corporation. *SSC-DC50A/50AP/54A/54AP/58AP Color Video Camera Operating Instructions*, 1998.

[7] Sony Corporation. *GV-D1000/D1000E Digital Video Cassette Recorder Operating Instructions*, 2002.

[8] Horita Corporation. *BSG-50 PAL Blackburst, Sync, Audio Tone Generator User Manual*, 1998.

[9] Mark of the Unicorn Inc. *MIDI Timepiece AV-USB User's Guide for Windows.*

[10] Horita Corporation. *VG-50 VITC Time Code Generator/LTC-VITC Translator User Manual*, 1994.