Lang Resources & Evaluation (2007) 41:389–407
DOI 10.1007/s10579-007-9054-4

The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms

Djamel Mostefa · Nicolas Moreau · Khalid Choukri · Gerasimos Potamianos · Stephen M. Chu · Ambrish Tyagi · Josep R. Casas · Jordi Turmo · Luca Cristoforetti · Francesco Tobia · Aristodemos Pnevmatikakis · Vassilis Mylonakis · Fotios Talantzis · Susanne Burger · Rainer Stiefelhagen · Keni Bernardin · Cedrick Rochet

Published online: 16 January 2008

Abstract The analysis of lectures and meetings inside smart rooms has recently attracted much interest in the literature, being the focus of international projects and technology evaluations. A key enabler for progress in this area is the availability of

Ambrish Tyagi has contributed to this work during two summer internships with the IBM T.J. Watson Research Center.

D. Mostefa (✉) · N. Moreau · K. Choukri
Evaluations and Language Resources Distribution Agency (ELDA),
55–57 rue Brillat Savarin, 75013 Paris, France
e-mail: mostefa@elda.org
URL: http://www.elda.org

N. Moreau
e-mail: moreau@elda.org

K. Choukri
e-mail: choukri@elda.org

G. Potamianos · S. M. Chu · A. Tyagi
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
URL: http://www.ait.gr

G. Potamianos
e-mail: gpotam@us.ibm.com

S. M. Chu
e-mail: schu@us.ibm.com

Present Address:
A. Tyagi
Department of Computer Science and Engineering, The Ohio State University,
Columbus, OH, USA

J. R. Casas · J. Turmo
Universitat Politècnica de Catalunya, Barcelona, Spain

J. R. Casas
e-mail: josep@gps.tsc.upc.edu

appropriate multimodal and multi-sensory corpora, annotated with rich human activity information during lectures and meetings. This paper is devoted to exactly such a corpus, developed in the framework of the European project CHIL, "Computers in the Human Interaction Loop". The resulting data set has the potential to drastically advance the state-of-the-art, by providing numerous synchronized audio and video streams of real lectures and meetings, captured in multiple recording sites over the past 4 years. It particularly overcomes typical shortcomings of other existing databases that may contain limited sensory or monomodal data, exhibit constrained human behavior and interaction patterns, or lack data variability. The CHIL corpus is accompanied by rich manual annotations of both its audio and visual modalities. These provide a detailed multi-channel verbatim orthographic transcription that includes speaker turns and identities, acoustic condition information, and named entities, as well as video labels in multiple camera views that provide multi-person 3D head and 2D facial feature location information. Over the past 3 years, the corpus has been crucial to the evaluation of a multitude of audiovisual perception technologies for human activity analysis in lecture and meeting scenarios, demonstrating its utility during internal

J. Turmo
e-mail: turmo@lsi.upc.edu

L. Cristoforetti · F. Tobia
ITC-IRST, Via Sommarive 18, 38050 Povo, Italy

L. Cristoforetti
e-mail: cristofo@itc.it

F. Tobia
e-mail: tobia@itc.it

A. Pnevmatikakis · V. Mylonakis · F. Talantzis
Athens Information Technology, Markopoulou Ave, 19002 Peania, Greece

A. Pnevmatikakis
e-mail: apne@ait.edu.gr

V. Mylonakis
e-mail: vmil@ait.edu.gr

F. Talantzis
e-mail: fota@ait.edu.gr

S. Burger
Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: sburger@cs.cmu.edu

R. Stiefelhagen · K. Bernardin · C. Rochet
Interactive Systems Labs, Universität Karlsruhe (TH), Karlsruhe, Germany

R. Stiefelhagen
e-mail: stiefel@ira.uka.de

K. Bernardin
e-mail: keni@ira.uka.de

C. Rochet
e-mail: crochet@ira.uka.de

evaluations of the CHIL consortium, as well as at the recent international CLEAR and Rich Transcription evaluations. The CHIL corpus is publicly available to the research community.

Keywords Mutlimodal · Corpus · Annotation · Evaluation · Audio · Video

1 Introduction

Interactive lectures and meetings play a significant role in human collaborative activities. Not surprisingly, analysis of interaction in these domains has attracted significant interest in the literature, being the central theme of a number of research efforts and international projects, most recently CHIL, "Computers in the Human Interaction Loop" (CHIL website), AMI, "Augmented Multi-party Interaction" (AMI website), the US National Institute of Standards and Technology (NIST) Smartspace effort (NIST smartspace), and a partial focus in the "Video Analysis and Content Extraction" (VACE) program (VACE website) and project CALO, "Cognitive Assistant that Learns and Organizes" (CALO website), among others. Of particular interest in some of these recent efforts is the case where the interaction happens inside smart rooms, equipped with multiple audio and visual sensors. Based on the resulting captured data, the goal is to detect, classify, and understand human activity in the space, addressing the basic questions about the "who", "where", "what", "when", and "how" of the interaction.

Achieving these objectives is the main focus of the European project CHIL, funded under the 6th Framework Programme. CHIL is a 3.5 years research effort with the participation of 15 partner sites from nine countries under the joint coordination of the Fraunhofer Institut für Informations- und Datenverarbeitung (IITB) and the Interactive Systems Labs (UKA) of the University of Karlsruhe, Germany (CHIL website). CHIL is driven by the desire to alter the traditional human computer interaction paradigm, advocating a new approach to provide more supportive and less burdensome computing and communication services to assist and facilitate human to human interaction during lectures and meetings. In the CHIL vision, computers fade into the background, reduced to "discreet" observers of human activity through the use of far-field sensors that allow the CHIL computing environment to detect, classify, understand, learn, and adapt to human activity. Central to this goal is the development of perception technologies that process the available multimodal and multi-sensory signals to robustly track position, recognize identity, process verbal communication information, and classify implicit human communication activity, such as emotional state, pose, focus of attention, and gestures of the participants. For such an effort to be successful, it needs to be accompanied by rigorous evaluation of the developed technologies, to allow performance benchmarking and a better understanding of possible limitations and challenging conditions. Clearly, a key enabler is the availability of appropriate corpora, annotated with necessary information, and accompanied with a suitable evaluation paradigm.

In this paper, we present our work over the past 4 years in acquiring such a database to allow development and evaluation of audiovisual perception technologies inside smart rooms, within the framework of the CHIL project. The resulting CHIL corpus has the potential to drastically advance the state-of-the-art in the area by providing numerous synchronized audio and video streams of 86 real lectures and meetings, captured in five recording sites. Although not the first corpus to address the meeting or lecture scenarios, it significantly overcomes deficiencies of other existing data sets. Indeed, the vast majority of such publicly available corpora focuses on the audio modality alone, mainly aiming at speech technology development, such as the ICSI (Janin et al. 2003) and ISL (Burger et al. 2002) meeting data sets. Other corpora recently collected by NIST (NIST smartspace) and the AMI project (AMI website) exhibit many similarities to the CHIL corpus, providing multimodal and multi-channel data sets inside smart rooms. However, they are either limited to a single data collection site or contain scripted and somewhat constrained, static interaction among the meeting participants. Therefore, to our knowledge, the CHIL corpus is the only data set that provides multimodal, multi-sensory recordings of realistic human behavior and interaction in lecture and meeting scenarios, with desirable data variability due to the numerous recording sites and sessions. These are key attributes that allow researchers to evaluate algorithmic approaches on real data, breaking the toy-problem barrier, to test generalization to mismatched data, and to experiment with channel and modality fusion schemes.

The CHIL corpus is accompanied by rich manual annotations of both its audio and visual modalities. In particular, it contains a detailed multi-channel verbatim orthographic transcription of the audio modality that includes speaker turns and identities, acoustic condition information, and named entities for part of the corpus. Furthermore, video labels provide multi-person head locations in the 3D space, as well as information about the 2D face bounding boxes and facial feature locations visible in the camera views. In addition, head-pose information is provided for part of the corpus.

Over the past 3 years, the CHIL corpus has been the corner-stone in the evaluation of a multitude of audiovisual perception technologies for human activity analysis during lectures and meetings. Such include person localization and tracking technologies, person identification, face recognition, speaker identification, gesture recognition, conversational large-vocabulary continuous speech recognition, acoustic scene analysis, emotion identification, topic identification, head-pose estimation, focus-of-attention analysis, question answering, and summarization. In more detail, the corpus has first been used in CHIL consortium internal technology evaluations (June 2004 and January 2005), followed by more recent international evaluation efforts. In particular, the corpus has been the main data set in the "CLassification of Events, Activities, and Relationships" (CLEAR) evaluation (Stiefelhagen and Garofolo 2006, CLEAR website) during the springs of 2006 and 2007. Furthermore, the corpus multi-channel audio modality has been part of the speech technology evaluations within the Rich Transcription (RT) Meeting Recognition evaluations organized by NIST (RT website) during the springs of 2005, 2006 and 2007, and will be used in a pilot track in the CrossLanguage Evaluation Forum CLEF 2007 (CLEF website). Utilization of the CHIL data set in these high-profile evaluation activities demonstrates the state-of-the-art nature of the corpus, and its contribution to advanced

perception technology development, further enhanced by the numerous papers resulting from these evaluations. The CHIL corpus is publicly available to the community through the language resources catalog (ELRA's Catalog) of the European Language Resources Association (ELRA).

The remainder of the paper is organized as follows. Section 2 provides an overview of the CHIL corpus, including a description of the lecture and meeting scenarios, as well as the characteristics of the smart room recording sites and resulting data. Section 3 is devoted to the corpus annotations concerning both its audio and visual modalities. This is followed by Sect. 4 that gives a brief overview of the evaluation paradigms for the various perception and content extraction technologies addressed by the CHIL corpus. Finally, Sect. 5 discusses corpus dissemination and Sect. 6 concludes the paper.

2 Corpus overview: data collection setup and scenarios

As already discussed in the Introduction, the CHIL corpus consists of multi-sensory audiovisual recordings inside smart rooms. The corpus has been collected over the past few years in five different recording sites and contains data of two types of human interaction scenarios: lectures and small meetings. In the following, we expand on these issues and provide a detailed overview of the corpus. We also describe the quality standard defined by the CHIL consortium in order to improve the data collection process.

2.1 Data collection setup

Five smart rooms have been set up as part of the CHIL project, and have been utilized in the data collection efforts. These rooms are located at the following partner sites: The Research and Education Society in Information Technologies at Athens Information Technology, Athens, Greece (AIT); the IBM T.J. Watson Research Center, Yorktown Heights, USA (IBM); the Centro per la ricerca scientifica e tecnologica at the Instituto Trentino di Cultura, Trento, Italy (ITC-IRST); the Interactive Systems Labs of the Universität Karlsruhe, Germany (UKA); and the Universitat Politècnica de Catalunya, Barcelona, Spain (UPC).

These five smart rooms are medium-size meeting or conference rooms with a number of audio and video sensors installed, and with supporting computing infrastructure. The multitude of recording sites provides desirable variability in the CHIL corpus, since the smart rooms obviously differ from each other in their size, layout, acoustic and visual environment (noise, lighting characteristics), as well as sensor properties (location, type)—see also Fig. 1. Nevertheless, it was crucial to produce a certain degree of homogeneity across sites to facilitate technology development and evaluations, which is why a minimum common hardware and software setup has been specified concerning the recording sensors and resulting data formats. All five sites comply with these minimum requirements, but often contain additional sensors. A minimal setup consists of:

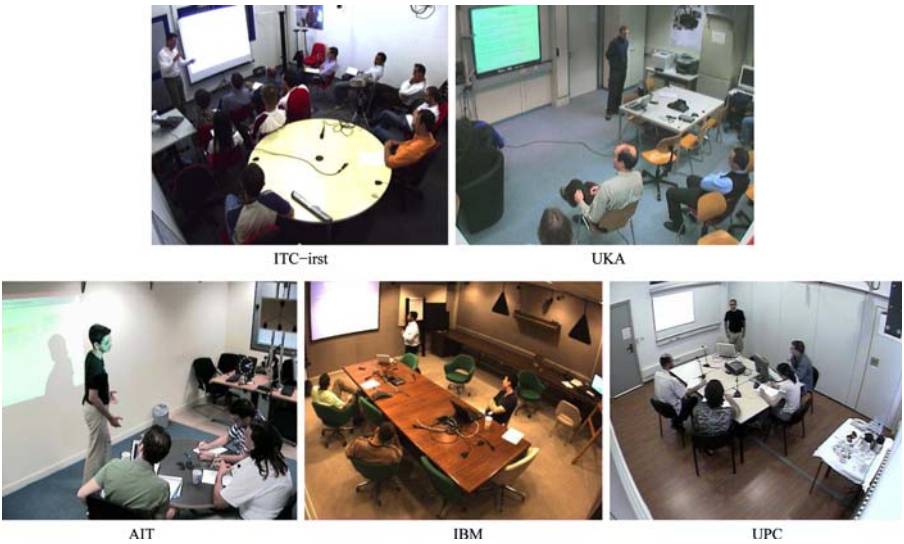

Fig. 1 Example camera views recorded at the five CHIL smart rooms during lectures (upper row) and small meetings (lower row)

– A set of common audio sensors, namely:
  - A 64-channel linear microphone array;
  - Three 4-channel T-shaped microphone clusters;
  - Three table-top microphones;
  - Close-talking microphones worn by the lecturer and each of the meeting participants.
– A set of common video sensors, that include:
  - Four fixed cameras located at the room corners;
  - One fixed, wide-angle panoramic camera located under the room ceiling;
  - One active pan-tilt-zoom camera.

This set is accompanied by a network of computers to capture the sensory data, mostly through dedicated data links, with data synchronization realized in a variety of ways. A schematic diagram of such a room including its sensors is depicted in Fig. 2. Additional details of the setup and the recorded data formats are given next.

2.1.1 Audio sensor setup

Each smart room contains a minimum of 88 microphones that capture both close-talking and far-field acoustic data. In particular, for far-field audio recording, there exists at least one 64-channel linear microphone array, namely the Mark III array developed by NIST (NIST MarkIII), placed on the smart room wall opposite to the speaker area. Such a sensor allows audio beam forming for speech recognition and

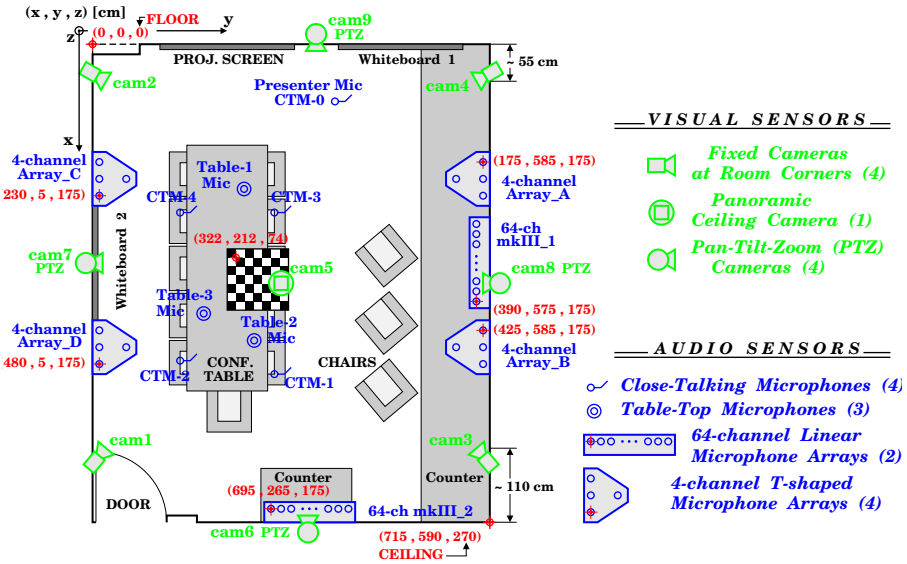

Fig. 2 Schematic diagram of the IBM smart room, one of the five installations used for recording the CHIL corpus. The room is approximately 7 × 6 × 3 × m^3 in size and contains nine cameras and 152 microphones for data collection

speaker localization. The microphone array is accompanied by at least three additional microphone clusters located on the room walls, each consisting of four microphones organized in an inverted "T" formation of known geometry to allow far-field acoustic speaker localization. Additional far-field audio is collected by at least three table-top microphones. The latter are positioned on the meeting table, but their exact placement is not fixed. As a contrast to the far-field audio data, close-talking microphones are used to record the lecture presenter and, in the case of small meeting recordings, all the meeting participants. At least one of these microphones is wireless, to allow free movement of the presenter. Slight variations of this setup can be found among the five recording sites. For example, the IBM smart room contains two NIST Mark III arrays, whereas the ITC room has seven T-shaped arrays.

For audio data capture, all microphones not belonging to the NIST Mark III are connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers. The pre-amplifier outputs are sampled at 44.1 kHz and 24 bits per sample, and are recorded to a computer in WAV format via an RME Hammerfall HDSP9652 I/O card. The 64-channel NIST Mark III data are similarly sampled and recorded in SPHERE format, but are fed into a recording computer via an ethernet connection in the form of multiplexed IP packets.

2.1.2 Video sensor setup

The video data is captured by five fixed cameras. Four of them are mounted close to the corners of the room, by the ceiling, with significantly overlapping and wide-angle fields-of-view. These are set in such a fashion, so that any person in the room is always

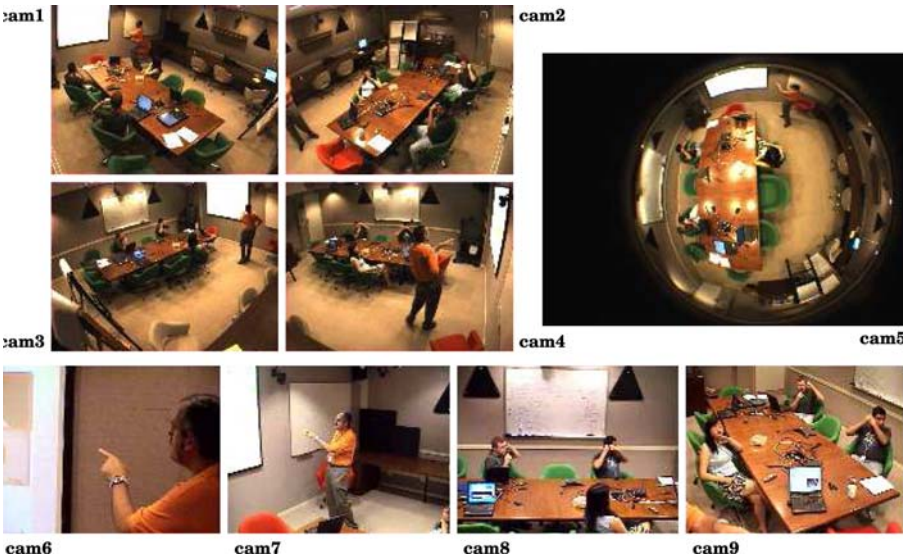

Fig. 3 Sample synchronous images captured at the IBM smart room during an interactive seminar (meeting). The five fixed camera views are depicted in the upper rows, pan-tilt-zoom camera views are shown in the lower row (see also Fig. 2)

visible by at least two cameras. The fifth camera is mounted on the ceiling, facing top-down, and uses a fish-eye lens to cover the entire room. The type of cameras installed varies among the sites, being either firewire or analog, providing images in resolutions ranging from 640×480 to 1024×768 pixels, and frame rates from 15 to 30 fps. All fixed cameras are calibrated with respect to a reference coordinate frame, with both extrinsic and intrinsic information provided in the corpus. In addition to the fixed cameras, at least one active pan-tilt-zoom (PTZ) camera is available in all five smart room setups. Its purpose is to provide close-up views of the presenter during lectures or meetings. There exist significant differences in the PTZ camera setups among the recording sites: the number of cameras used, their type (analog, digital), their control medium (serial, network), as well as the control mechanism (human operator, automatic). An example of smart room camera views is depicted in Fig. 3.

For data capture, a number of dedicated computers are used, with all video streams saved as sequences of JPEG-compressed images. This allows easy non-linear access to the frames, as well as exact absolute time stamping. It is also worth mentioning that most meeting recordings are accompanied by brief video sequences that contain empty room images, captured immediately preceding the entry of all participants. These are provided to assist background modeling in video processing algorithms.

2.2 Quality standards

In order to have every site producing the same quality of data, CHIL developed internally a standard of quality for all the sensors.

2.2.1 Video quality standard

Each site followed the recommendation of four angle cameras and a central ceiling mounted fish eye camera. The minimum frame rate was set to 15 frames per seconds (fps). The data streams were saved as sequences of JPEG images in a fixed name standard: seq xxxxx.jpg with xxxxx the number of the frame. A specific file called seq.index contained the table of correspondence between the frame and its associated time stamp. A file called seq.ini contained all the camera related information.

The maximum desynchronization between the five cameras for the entire length of a recording was set to 200 ms.

This was measured by introducing at the beginning and end of each recording a distinct and well observable audio-visual signal. The decision on how to realize this was left to the recording site but a movie studio-type clap was suggested. This was also a good way of testing the synchronization between the audio and video channels.

2.2.2 Microphone array quality standard

Each site was equipped with at least one fully functional Mark III microphone array version 2. The version 2 was developed in collaboration with NIST. It generates 64 channels of audio, captured at 44 KHz and 24 bits of resolution. For each recording, the channel 4 was extracted. A specific file called timestamps.ini was created to store the time stamp of an eventual packet loss. The maximum desynchronization due to packet loss during one recording was fixed to 200 ms. If more occurred, the recording had to be remade.

2.2.3 Hammerfall quality standard

Each site was equipped with at least 20 microphones for synchronous capture of audio. The former correspond to at least three T-shaped microphone arrays, each having four microphones, located on the walls. The remaining channels are from table-top microphones located on the conference table and close-talking ones. Just as for the MarkIII microphone array, a specific file called timestamps.ini was created to store the time stamp of an eventual packet loss. The maximum desynchronization due to packet loss during one recording was fixed to 50 ms. If more occurred, the recording had to be remade.

2.2.4 Additional information

To have every site providing the same information in a structured manner, a specific info directory was included in each recording. It contains a calibration directory with 10 pictures per camera and their calibration results and a background directory with pictures of the background before the meeting, when the room is empty.

A seminar datasheet was also required for each recording. It mainly contains information about the attendees: photo with identity tags, microphones corresponding to each attendee, etc. The presentation slides were also required. All this information was meant to make the transcription and annotation work easier and more reliable.

2.3 Data collection scenarios

Two types of interaction scenarios constitute the focus of the CHIL corpus: lectures and meetings. In both cases, a presenter gives a seminar in front of an audience, but the two scenarios differ significantly in the degree of interactivity between the audience and the presenter, as well as the number of participants. The seminar topics are quite technical in nature, spanning the areas of audio and visual perception technologies, but also biology, finance, etc. The language used is English, however most subjects exhibit strong non-native accents, such as Italian, German, Greek, Spanish, Indian, Chinese, etc. More information about the two seminar classes follows.

2.3.1 Non-interactive seminars (lectures)

In this scenario, the presenter talks in front of an audience of typically 10–20 people having little interaction in the form of a few question–answering turns, mostly towards the end of the presentation. As a result, the audience region is quite cluttered and of little activity and interest. The focus in lecture analysis lies therefore on the presenter. As a consequence, only the presenter has been annotated in the lecture part of the CHIL corpus (see Sect. 3) and is the subject of interest in the associated evaluation tasks (see Sect. 4). Examples of non-interactive seminars are depicted in the upper row of Fig. 1. Such data have been recorded at UKA during 2003, 2004, and 2005, and at ITC in 2005. A total of 46 lectures are part of the CHIL corpus (see also Table 1), most of which are between 40 and 60 min long.

2.3.2 Interactive seminars (meetings)

In this scenario, the audience is small, between three and five people, and the attendees mostly sit around a table, all wearing close-talking microphones. There exists significant interaction between the presenter and the audience, with numerous questions and often a brief discussion among meeting participants.

Typically, such scenarios include the following events:

- participants enter or leave the room,
- some attendees stand up and go to the whiteboard,
- discussions among the attendees,
- participants stand up for a short coffee break,

Table 1 Details of the 86 collected lectures/non interactive seminars (upper table part) and meetings/interactive seminars (lower part) that comprise the CHIL corpus

Site	# Seminars	Year	Type	Evaluations
UKA	12	2003/2004	Lectures	CHIL Internal Evals., CLEF07
UKA	29	2004/2005	Lectures	CLEAR06, RT05s, RT06s, CLEF07
ITC	5	2005	Lectures	CLEAR06, RT06s
UKA	5	2006	Meetings	CLEAR07, RT07s
ITC	5	2006	Meetings	CLEAR07, RT07s
AIT	5	2005	Meetings	CLEAR06, RT06s
AIT	5	2006	Meetings	CLEAR07, RT07s
IBM	5	2005	Meetings	CLEAR06, RT06s
IBM	5	2006	Meetings	CLEAR07, RT07s
UPC	5	2005	Meetings	CLEAR06, RT06s
UPC	5	2006	Meetings	CLEAR07, RT07s

The table depicts the recording site, year, type, and number of collections, as well as the evaluations where the data were used

– during and after the presentation there are questions from the attendees with answers from the presenter.

In addition, a significant number of acoustic events is generated to allow more meaningful evaluation of the corresponding technology:

– sounds when opening and closing the door,
– interruptions of the meeting due to ringing mobile phones,
– attendees coughing and laughing,
– attendees pouring coffee in their cup and puting it on the table,
– attendees playing with their keys,
– keyboard typing, chair moving, etc.

Clearly, in such a scenario all participants are of interest to meeting analysis, therefore the CHIL corpus provides annotations for all (see Sect. 3). Examples of interactive seminars are depicted in the lower row of Fig. 1. Such data have been recorded by AIT, IBM and UPC in 2005, as well as all five sites during 2006. A total of 40 meetings are part of the CHIL corpus, most of which are approximately 30 min in duration. Table 1 summarizes the recorded data sets, when and where they were recorded, and in which evaluation they were used (For further details on evaluations, see Sect. 4).

3 CHIL corpus annotations

For the collected data to be useful in technology evaluation and development, it is crucial that the corpus is accompanied by appropriate annotations. The CHIL consortium has devoted significant effort in identifying useful and efficient

annotation schemes for the CHIL corpus and providing appropriate labels in support of these activities. As a result, the data set contains a rich set of annotations in multiple channels of both audio and visual modalities. Details are discussed next.

3.1 Audio channel annotations

Data recording in the CHIL smart room results in multiple audio files containing signals recorded by close-talking microphones (near-field condition), table-top microphones, T-shaped clusters, and the Mark III microphone array (far-field condition), in parallel. The recorded speech as well as environmental acoustic events were carefully segmented and annotated by human transcribers at two locations, the European Language Resources Distribution Agency (ELDA) and the interACT Center at Carnegie Mellon University (CMU).

3.1.1 Orthographic transcriptions

Transcriptions were done by native English speakers. Detailed transcription guidelines were given to the transcribers to define common rules of annotations. The manual transcription process started by transcribing the speaker contributions of all recorded near-field channels on orthographic word level, including the typical speaker-produced noises such as laughter and filled pauses. The start and end of the contributions were manually segmented. The transcription of the near-field condition was then compared to one of the far-field channels. Non-audible events were removed and details recorded by only the far-field sensors were added. The transcription factor was quite important. In average, it took 30 h to transcribe 1 h of signal from one channel.

In the case of the non-interactive seminars (lectures), the near-field condition needed to be transcribed only from the recording of the presenter, since he/she was the only one with a close-talking microphone. Human annotators used the Transcriber tool (Transcriber). In a second step, the transcription of the near-field condition was adapted to a far-field recording, typically one of the channels recorded by the Mark III microphone array. Environmental noises were added, as well as contributions from the audience. Non-audible parts, which were only heard during the near-field transcription, were removed.

An example of the resulting transcripts is depicted in Fig. 4. The left window shows the near-field transcription of several turns in a lecture, while the right window depicts the transcription of the far-field condition of the same lecture. It can be seen that speaker "UPC_002", a participant sitting in the audience, was not audible in the near-field condition where his utterances are tagged as "inaudible speech (is)". However, the same utterances were clearly understood in the far-field condition and, therefore, are included in the transcription.

In contrast to the non-interactive seminars, all participants of the interactive seminars (meetings) wore individual close talking microphones. These conversations contain overlapping speech and discursive turns reacting to each other

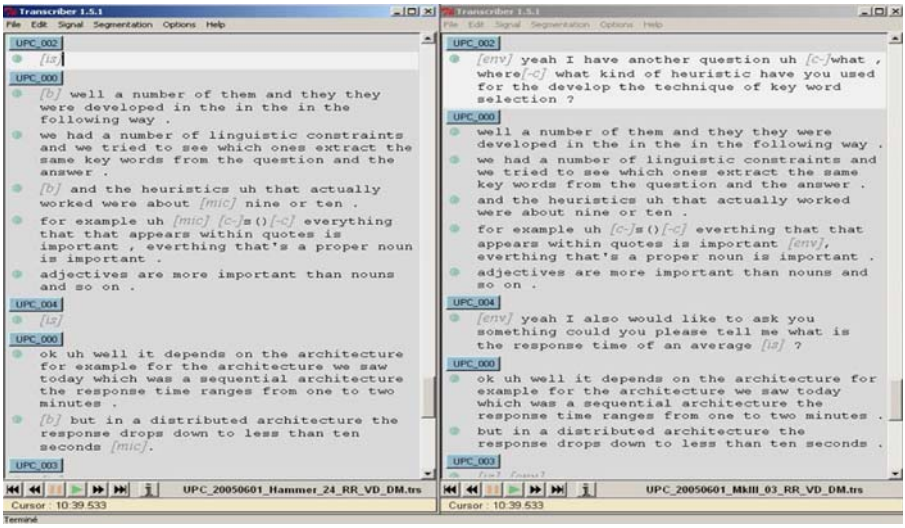

Fig. 4 Transcription of a lecture segment using the Transcriber tool. Near-field transcription (left side) versus far-field transcription (right side) are shown

(question/answer, argument/counter argument), similar to a multi-party meeting conversation. To display all recorded speech signals in a multi-track parallel view, transcribers worked with TransEdit, a tool originally developed for the transcription of multi-party meetings. The near-field transcription was again compared to the far-field condition. The far-field transcription of an interactive seminar contains only the transcription of the speech; non-audible parts were removed, or marked as difficult to understand. Background speaker contributions were added. Many of the segmentation boundaries needed to be adjusted because reverberation in the far-field recordings resulted in shifted segment boundaries.

3.1.2 Annotation of acoustic events

Following the orthographic transcription of close-talking and far-field audio, a third step was performed for annotating environmental acoustic events. Such annotations were used in support of the "acoustic event detection and classification" task in the CLEAR evaluations.

Acoustic events describe all audible events in a recording. Accordingly, SPEECH is here also considered an acoustic event but is only broadly labeled as SPEECH, not transcribed in single words. Beside SPEECH, the set of labels for acoustic events consists of DOOR SLAM, STEP, CHAIR MOVING, CUP JINGLE, APPLAUSE, LAUGH, KEY JINGLE, COUGH, KEYBOARD TYPING, PHONE RINGING, MUSIC, KNOCK (door, table), PAPER WRAPPING, and UNKNOWN.

The annotation of acoustic events was carried out as an independent additional labeling process using the Annotation Graph Tool Kit (AGTK). Unlike Transcriber, AGTK enables the annotation of multiple overlapping events (AGTK website).

Acoustic events were labeled on two different types of data sets: acoustic events occurring in the CHIL lecture and meeting corpus and recordings of artificially produced events. The first set of data was labeled listening to the fourth channel of the Mark III microphone array. The artificially produced acoustic events were recorded in two data sets in the ITC and UPC smart rooms, and they contain isolated acoustic events collected in a quiet environment with no temporal overlap.

3.2 Video channel annotations

3.2.1 Facial features and head location information

Video annotations were manually generated using an ad-hoc tool provided by the University of Karlsruhe and modified by ELDA. The tool allows displaying one picture every second, in sequence, for all camera views. To generate labels, the annotator performs a number of clicks on the head region of the persons of interest, i.e., the lecturer only in the non-interactive seminar (lecture) scenario, but all participants in the interactive seminar (meeting) scenario. In particular, the annotator first clicks on the head centroid (e.g., the estimated center of the person's head), followed by the left eye, the right eye, and the nose bridge (if visible). In addition, the annotator delimits the person's face with a bounding box. The 2D coordinates of the marked points within the camera plane are saved to the corresponding label file. This allows the computation of the 3D head location of the persons of interest inside the room, based on camera calibration information. Figure 5 depicts an example of video labels, produced by this process. It shows the head centroid (white), the left eye (blue), the nose bridge (red), the right eye (green), and the face bounding box.

3.2.2 Head pose annotations

In addition to 2D face and 3D head location information, part of the lecture recordings were also labeled with gross information about the lecturer's head pose.

Fig. 5 Example of video annotations for an interactive seminar in the UPC smart room. Face bounding boxes and facial feature annotations are depicted for two camera views

In particular, only eight head orientation classes were annotated, deemed to be a feasible task for human annotators, given the low-resolution captured views of the lecturer's head. The head orientation label corresponded to one of eight discrete orientation classes, ranging from a 0° to a 315° angle, with an increment of 45°. Overall, 19 lecture videos were annotated with such information. These videos were used in the CLEAR head-pose technology evaluation.

3.3 Validation procedures

The video annotations were validated internally. After being produced by human annotators, each annotation file was automatically scanned using a tool developed by ELDA. This tool detects most of the annotation errors that can occur: inversion of right and left eyes, missing labels, etc. During a second validation pass, a human operator checked and corrected manually the video labels. The error listings produced by the automatic scanning tool helped in this task. It was ensured that the person who checked a given seminar was different from the one who initially labeled it.

In the same way, each orthographic transcription was validated by a human transcriber, different from the one who produced it. A final pass was performed where all the data were reviewed by one person who used semi-automatic methods (spellchecker, lexicon, list of proper names,...) to check and correct the data.

A further cross-validation check of video labels (at UKA) and audio transcriptions (between ELDA and CMU) was done. A few annotations were examined at random, to check if they were correct.

4 Technology evaluations using the CHIL corpus

As mentioned before, the CHIL corpus was designed to support the development and evaluation of multimodal technologies for the perception of humans, their interaction and activities in realistic lectures and meetings. Both the nature of the data (multimodal, multi-sensory) and the realistic, unconstrained human interaction behavior in recordings challenge the most advanced technologies targeting human activity analysis in indoor scenarios. In the CHIL corpus, perception technologies are faced with multiple concurrent variability in the audio-visual data. Far-field microphones, harsh acoustic environments, with noise and reverberation, and the need for automatic segmentation are demanding situations for acoustic and speech analysis technologies, whereas video technologies have to deal with low resolution images, uncontrolled illumination, real, changing backgrounds, varied environments, with different geometries and lighting, and the uncooperative attitude of the user (e.g. not facing the cameras). Most of these variations are often present together at the same time.

Different parts of the CHIL corpus have been used in a number of evaluations of perception technologies (see also Table 1). The first such evaluations were conducted internally within the CHIL consortium, first as a dry-run in June 2004,

followed by the January 2005 CHIL evaluation. Subsequently, the CHIL technology evaluations have been conducted mainly within two open international evaluation workshops: the newly created Classification of Events, Activities and Relationships—CLEAR—evaluation (CLEAR website), which was jointly organized by UKA and NIST in 2006 and 2007, and the NIST Rich Transcription (RT) Meeting evaluation (RT website). While CLEAR focuses on the analysis of technologies for the perception of people, their activities and interaction, RT focuses on technologies for language transcription. Furthermore, in 2007, a part of the CHIL corpus will be used to evaluate Question Answering technologies within CLEF 2007, in which a new task on Question Answering will be added (CLEF website).

The technology evaluations addressed in these CHIL internal and international evaluation workshop included the following evaluation tasks:

- *Person tracking*: here, the task was to track one or all people in the scene (3D) or in all images (2D). Subtasks included acoustic, visual and audio-visual tracking.
- *Person identification*: here, the task was to identify all people in the scenario. Subtasks included acoustic, visual, and audio-visual identification, as well as different lenghts of audio-visual segments for training and testing.
- *Head pose estimation*: here, the task was to estimate the 3D head orientation of a selected person (e.g. the lecturer) using all camera views.
- *Speech activity detection and speaker diarization*: here, the task was to correctly detect speech segments. In speaker diarization, furthermore the identities of speakers had to be determined (the "who spoke when" problem).
- *Automatic speech recognition (ASR)*: here, the task was to produce a transcription of the speakers' speech. Subtasks included different microphone conditions, such as ASR from close-talking (CTM), table-top (TT) or far-field microphones (FF).
- *Acoustic event detection*: here, the task was to detect and identify a number of acoustic events in meeting or lecture data.
- *Question answering (QA) and summarization (SA)*: these tasks addressed technologies to provide answers to user questions extracted from speech transcripts. Answers can be from exact and specific facts, such as names of persons and organizations, to summaries.

Table 2 summarizes the tasks evaluated using the CHIL corpus, and which evaluations they have been conducted in. Further details of the evaluation tasks, including detailed evaluation procedures and metrics can be found in (Mostefa et al. 2005; Stiefelhagen et al. 2007).

5 Evaluation packages and dissemination

The CHIL corpus is publicly available to the academic and industrial communities as a set of "evaluation packages" through the ELRA General Catalog (ELRA's Catalog).

An evaluation package consists of full documentation (including definition and description of the evaluation methodologies, protocols, and metrics), along with the

Table 2 Overview of the tasks and evaluation workshops, which used part of the CHIL corpus

Tasks	Evaluation
Person tracking (2D, 3D, A, V, AV)	CHIL internal, CLEAR'06, CLEAR'07
Person identification (A, V, AV)	CHIL internal, CLEAR'06, CLEAR'07
Head pose estimation	CHIL internal, CLEAR'06, CLEAR'07
Acoustic event detection	CHIL internal, CLEAR'06, CLEAR'07
Speech recognition (CT, FF, TT mics)	RT'05, RT'06, RT'07
Speech activity detection	CHIL internal, RT'05, RT'06, RT'07
Speaker diarization	RT'07
Question answering	CHIL internal, CLEF'07
Summarization	CHIL internal

data sets and software scoring tools, necessary to evaluate developed systems for a given technology. Such a package therefore enables external participants to benchmark their systems and compare results to those obtained during the official evaluation.

The CHIL Evaluation Packages consist of the following:

- A document describing in detail the content of the package, as well as the corresponding evaluation (tasks, metrics, participants, results, etc.),
- The raw audio recordings of the seminars (Hammerfall, close talking microphones and microphone array channels),
- The raw video recordings of the seminars (streams of the four corner cameras and ceiling camera),
- The video annotations and audio transcriptions of the seminars,
- Useful information about each seminar (attendees, slides, calibration information, background pictures),
- Additional databases specific to some evaluation tasks (head pose, pointing gestures, isolated acoustic events).

In addition, a range of specific data is provided for each evaluation task, allowing the package user to reproduce the evaluation in the same conditions and to compare his results with those of the participants:

- Documentation about the evaluation procedure (metrics, submission format, etc.),
- The input data, as received by the participants during the evaluation,
- The participants' submissions,
- The reference labels,
- The scoring tools,
- The participants' results.

So far, two evaluation package "suites" have been produced: The first is composed of evaluation packages stemming from the CHIL-internal evaluation that took place in January 2005 (Mostefa et al. 2006). A description of the data, tools,

and results is available in a public document (Mostefa et al. 2005). The second suite covers the CLEAR 2006 evaluation (Stiefelhagen et al. 2007). A third is planned to be released following completion of the CLEAR 2007 evaluation (see also Table 1).

6 Conclusions

In this paper, we have presented an overview of the CHIL corpus, a one-of-a-kind audiovisual database of lectures and meetings held inside smart rooms that are equipped with a multitude of sensors. The corpus has been collected as part of the CHIL project, aiming in developing and evaluating audiovisual perception technologies concerning human activity and interaction during lectures and meetings. The resulting data set has largely contributed to advancing the state-of-the-art in the field by providing numerous synchronized audio and video streams of 86 real lectures and meetings, captured in five recording sites over the past 4 years. Significant effort has been dedicated to accompany the recorded data with rich multi-channel annotations in both audio and visual modalities. The CHIL corpus has already been utilized in international evaluations and is publicly available to the research community. We therefore strongly believe that it represents an important contribution and a resource crucial to the development of robust perception technologies for the analysis of realistic human interaction.

Acknowledgments The work presented here was partly funded by the European Union under the integrated project CHIL, "Computers in the Human Interaction Loop" (Grant Number IST-506909).

References

AMI—Augmented Multiparty Interaction. http://www.amiproject.org

Burger, S., McLaren, V., & Yu, H. (2002). The ISL meeting corpus: The impact on meeting type on speech style. In *Proceedings of International Conference on Spoken Language Processing*, Denver, USA.

CALO—Cognitive Agent that Learns and Organizes. http://www.caloproject.sri.com/

CHIL—Computers in the Human Interaction Loop. http://www.chil.server.de

Classification of Events, Activities, and Relationships Evaluation and Workshop. http://www.clear-evaluation.org

ELRA Catalogue of Language Resources. http://www.catalog.elra.info

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China.

Mostefa, D., et al. (2005). Chil Public Deliverable D7.6: Exploitation material for CHIL evaluation campaign 1. http://www.chil.server.de/servlet/is/8063/

Mostefa, D., Garcia, M.-N., & Choukri, K. (2006). Evaluation of multimodal components within CHIL. In *Proceedings of the 5th International Language Resources and Evaluations Conference (LREC)*, Genoa, Italy.

Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., & Soundararajan, P. (2007). The CLEAR 2006 evaluation. In R. Stiefelhagen & J. Garofolo (Eds.), Multimodal Technologies for Perception of Humans. *Proceedings of the First International CLEAR Evaluation Workshop, CLEAR 2006*, number 4122 in Springer Lecture Notes in Computer Science, pp. 1–45.

Stiefelhagen, R., & Garofolo, J. (Eds). (2007). *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR'06*. Number 4122 in Lecture Notes in Computer Science, Springer.

The AGTK Annotation Tool. http://www.agtk.sourceforge.net
The CLEF Website. http://www.clef-campaign.org/
The NIST MarkIII Microphone Array. http://www.nist.gov/smartspace/cmaiii.html
The NIST Smart Space Project. http://www.nist.gov/smartspace/
The Rich Transcription 2006 Spring Meeting Recognition Evaluation Website. http://www.nist.gov/speech/tests/rt/rt2006/spring
The Transcriber Tool Home Page. http://www.trans.sourceforge.net
VACE—Video Analysis and Content Extraction. https://www.control.nist.gov/dto/twiki/bin/view/Main/WebHome