A multilingual corpus for rich audio-visual scene description in a meeting-room environment

Taras Butko
Universitat Politècnica de Catalunya
c/ Jordi Girona, 1-3, Campus Nord D5
Barcelona, 08034, Spain
+34 934011627
taras.butko@upc.edu

Climent Nadeu
Universitat Politècnica de Catalunya
c/ Jordi Girona, 1-3, Campus Nord D5
Barcelona, 08034, Spain
+34 934016438
climent.nadeu@upc.edu

Asuncion Moreno
Universitat Politècnica de Catalunya
c/ Jordi Girona, 1-3, Campus Nord D4
Barcelona, 08034, Spain
+34 934016437
asuncion.moreno@upc.edu

ABSTRACT
In this paper, we present a multilingual database specifically designed to develop technologies for rich audio-visual scene description in meeting-room environments. Part of that database includes the already existing CHIL audio-visual recordings, whose annotations have been extended. A relevant objective in the new recorded sessions was to include situations in which the semantic content can not be extracted from a single modality. The presented database, that includes five hours of rather spontaneously generated scientific presentations, was manually annotated using standard or previously reported annotation schemes, and will be publicly available for the research purposes.

Categories and Subject Descriptors
H.2.4 **[Database Management]**: Systems – *multimedia databases.*

General Terms
Measurement, Design, Reliability, Standardization, Languages.

Keywords
Database, multimodal, multilingual, audio-visual scene description

1. INTRODUCTION
Recently, describing, understanding and exploring interaction process among people, and people with the world, has attracted significant interest in the literature, being a great scientific challenge. It has been the main focus of research in several research projects like CHIL [2] and AMI [11]. Human interaction is a complex process that involves not only speech communication but also multiplicity of non-verbal cues including acoustic events, facial expressions, head gestures, gaze, body postures, focus of attention, prosodic cues etc. The automatic description of interactions between humans and environment can be useful for providing: implicit assistance to the people inside the room, context-aware and content-aware information requiring a minimum of human attention or interruptions, support for high-level analysis of the underlying acoustic scene, etc. Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments like meeting-rooms. In such environments a wide range of audiovisual perceptual technologies is needed, that can deliver the relevant information about what is going on in the scene based on the analysis of acoustic and visual sensors. This includes information about the number of people, their identities, activity, locations, postures, mood, gestures, body and head orientations, etc.

State-of-the-art technologies for a rich audio-visual scene description depend to a large extend on sufficient and appropriate sample data, often covering a particular domain, acoustic environment, recording channel or modality. Usually, the existing databases are contain only part of the whole variety of possible annotations, and they are not purposefully designed to force the joint and collaborative use of several technologies. This is the case for the multimodal database produced in the framework of the EU-funded CHIL project (2004-2007) [2], that corresponds to interactive seminars and presentations in English, recorded in smart-room environments, using cameras, and both far-field and close-talk microphones. The available annotations include only orthographic transcription of speech and other sounds, as well as the identity and position of the participant persons. Let us mention also a few other examples. For instance, the multimodal USC CreativeIT database [3] is intended for research on theoretical improvisation and human expressive behavior in dyadic interaction. Only motion parameters and emotional attributes are transcribed for each participant. The multimodal corpus for gesture expressivity analysis in [4] is oriented towards studying the use of hand emotional gestures in combination with facial expressions and speech. The French spontaneous conversations database [5] is intended for annotation of gesture, posture and gaze for sitting speakers. There are also more "natural" corpora like videolecture.net and ted.com, which include presentations or talks, and are large and freely available. However, the uncontrolled conditions of the recordings make difficult to use them for research purposes.

The objective of our work is creating a database specifically oriented to rich description of the underlying audio-visual scene in meeting-room environments, as if the purpose was to precisely describe with words the scene to both a deaf person and a person outside the room (or blind). As a starting point we chose the above mentioned CHIL database, and its existing annotation has been upgraded so that: 1) it can also be used with other speech technologies, and 2) it includes rich human activity information. Also, new multimodal recordings are performed whose design is specifically focused to that rich audio-visual description purpose. In our annotations we have used standard or previously reported

annotation schemes. For spatial annotation, the existing Spatial Role Labelling approach is used [6]. For emotion information, the EmotionML (Emotion Markup Language) standard [8] is applied. The result is a multimodal resource designed for research work on the various involved technologies, and in particular on the integration of their outputs for a multi-level based scene analysis. Three languages are used (English, Spanish and Catalan) and the total duration is 5 hours.

The work is being performed in the framework of the METANET4U European project, that aims at supporting language technology for European languages and multilingualism. The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them [9].

2. DATABASE

2.1 CHIL 2007 Evaluation Package

The CHIL 2007 Evaluation Package was produced within the CHIL Project [2] and includes scientific presentations given by students, faculty members or invited speakers in the field of multimodal interfaces and speech processing. The objective of this project was to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves [2].

The CHIL 2007 Evaluation Package consists of the set of audiovisual recordings of interactive seminars produced by AIT (Athens Information Technology), ITC (Instituto Trentino di Cultura), IBM, UKA (Universität Karlsruhe), and UPC (Universitat Politècnica de Catalunya) in their smart-rooms. Each seminar usually consists of a presentation of 20 to 30 minutes to a group of three to five attendees in a meeting room. During and after the presentation, there are questions from the attendees with answers from the presenter. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, discussion among the attendees, coffee breaks, etc. The number of people present in the recording was fixed to be between 3 and 7. There are 2 types of audio-visual documents presented in CHIL database: full sessions of about 20 minutes and excerpts of 5 min. The full sessions contain the meeting starting from its beginning until the end. The excerpts include short and the most prominent pieces of meeting where interactions, discussions and movements happen frequently. The evaluation package includes 1 full session and 8 excerpts from participant site.

2.2 Design and recording of the new sessions

In addition to the existing database described in sub-Section 2.1, two new sessions of approximately 30 min each have been recorded in Spanish and Catalan language, separately.

The general scheme of the recording sessions is similar to the one proposed in CHIL database. There are four participants in each of the sessions who interact among them in a natural and spontaneous way. During the recorded session the main presenter conducts the meeting and asks the other 3 people (experts) about their opinion related to the discussed topic. There is a coffee break in the middle of recording session, were the participants drink coffee, water and perform informal talks.

The main objective during the design of the sessions was oriented towards creating situations where there exists an ambiguity if only one modality is considered. Mainly, this corresponds to the cases when speech information is not sufficient to understand what people are intended to say. For instance, the presenter may not say explicitly which object or subject he is referring to but show it using hand gestures or by means of focus of attention. The recorded material is specifically designed for development of technologies that use all available information from different modalities and describe the scene in unambiguous way.

2.2.1 Recordings

The database was created using 24 T-shape microphones, 4 omni-directional microphones located on the table and 4 close-talk microphones (Figure 1). Regarding video, 5 cameras were employed: 3 cameras in the corners of the room (cam2, cam3, cam4,), one zenithal (cam5) and one lateral camera (cam7).

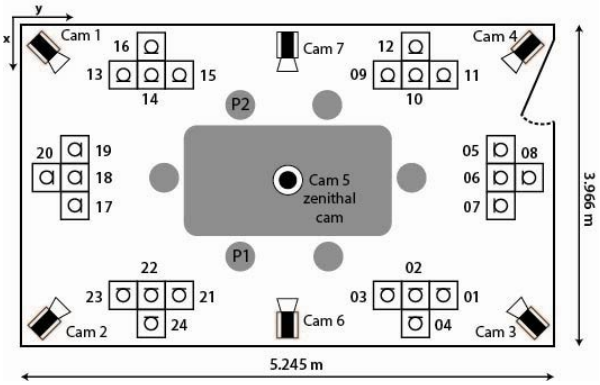

Figure 1. Top view of the UPC room.

The participants were asked to act in natural way, with the tendency to express emotions in rather natural way. They were asked to produce movements, make hand, head gestures etc. Six persons participated in recordings: 4 male speakers and 2 female, 5 native and 1 non-native speakers.

2.2.2 Creating the need of both modalities to avoid ambiguities

The scenarios of the new database were created in such a way that the information from a single modality is not sufficient to describe fully and in unambiguous way the interaction process in the room. In this case, the information from both audio and video is complementary, and a joint use of both modalities can resolve the ambiguities presented by a single modality.

One possible case of ambiguity in the audio modality is related to the use of deictic expressions in conversation. Deixis is a reference to an object or subject by means of an expression whose interpretation is relative to the (usually) extralinguistic context of the utterance; e.g. who is speaking, the time or place of speaking, the gestures of the speaker, the place in the discourse. We can distinguish 3 types of deixis produced by the participants of recordings:

1. Personal deixis (e.g. *me, you*)

2. Spatial deixis (e.g. *here, there*)

3. Temporal deixis (e.g. *now, then*)

Note that the information from the video modality is usually helpful to disambiguate either the personal or the spatial deixis, but not the temporal one. Consider the situation when the presenter says to the audience: "*We will record the video signal from **that** camera …*". This sentence doesn't contain the information about which camera is intended to be used, so a certain ambiguity in the audio modality exists. At the same time, the presenter may point to the concrete camera or gaze to it (Figure 2), so the video signal will contain the information missing in the audio signal. In the scenario of recordings people talk about objects in the room. For instance, there are several microphones at different positions of the room, 2 doors, multiple chairs around the table, etc. The participants were instructed to make as much use as possible of deixis when referring to these objects, but it was not indicated at specific places of the scripts.

(a)

(b)

Figure 2. Interaction between modalities: (a) pointing to the microphone cluster used in explanations (b) pointing to the door used in explanation (there are 2 doors in the room).

Apart from deictic expressions, the video modality is useful to recover missing oral answers to questions like: "Who will be the next presenter?" The answer "I am" may be missing, as just a hand or head gesture may be the answer. Again the participants were instructed to avoid explicit use of names by substituting them by pronouns where possible.

In Table 1 we present a short summary of possible cases when audio and video information can be fused to get the information missed by a single modality.

Table 1. Possible cases of complementarity of audio and video modalities.

Audio	Video	Result
Spatial deixis: This, that, it, there, here	Pointing gesture, focus of attention, position	Object
Person deixis: he/she, you, they	Pointing gesture, focus of attention,	Subject
Questions: who?, how many?, general question.	Hand gesture, head gesture	Missing answers to the question

3. ANNOTATION

The audio-visual recordings of the CHIL evaluation package are extended, and the existing annotation is completed. In Table 2 we summarize different annotation layers (tiers) together with the categories that are included to both the CHIL and the Metanet4u database. We consider four main annotation categories: audio, video, joint audio-video and text. From Table 2 one can observe that the existing CHIL annotation was significantly extended for video, joint audio-video and text categories. Compared with the AMI database annotations, the Metanet4u database annotation includes several new tiers: position of the participants, spatial roles, emotion information, and links between tiers. The annotation was performed by 2 annotators using the ELAN (EUDICO Linguistic Annotator) annotation tool [12], that allows to create, edit, visualize and search annotations for video and audio data. The text content of annotations is in Unicode and the transcription is stored in an XML format.

Table 2. Annotation layers in previously annotated CHIL and a new METANET database.

Annotation tiers	Categories	Database	
		CHIL	METANET
Audio signal			
Speech transcription	• word utterances • silence • vocalizations (e.g. <uh>, <uhm>, <Smack>, <B>)	√	√
Speaker labels	labels for each speaker	√	√
Acoustic events	12 non-speech sounds: cough, laugh, door slam, chair moving,	√	√
Video signal			
Activity classification	• presentation • discussion • informal talk (coffee break)		√
Movement	• off camera • sit • take notes • move		√

	• stand whiteboard		
Position of the participants	• 3-D coordinates of each participant	√	√
Focus of attention	• object or person		√
Hand gestures	• raising a hand • pointing gesture		√
Head gestures	• agreement signal • disagreement signal		√
Spatial role labeling (SRL)	• Trajector • Landmark • Motion • Spatial indicator		√
Audio-visual signal			
Emotion information	• neutral, • emphasized • negative (frustrated) • enthusiastic		√
Links between tiers	explicit links between deictic expressions from orthographic transcription and the corresponding transcription in video layer		√
Text			
Named entities	Unique identifiers of entities: • person, locations, organizations • dates times and durations • quantities, measures, percents, cardinal numbers • technical expressions		√
Topics	• top-level topics • sub topics. • functional descriptions		√

3.1 Audio signal annotation

The orthographic transcriptions for the new recordings in Spanish and Catalan were done by native speakers. Detailed transcription guidelines were given to the transcribers to define common rules of annotations as described in [1]. Besides, 12 classes of acoustic events which naturally occur in meeting-room environments [11] have been annotated: "Door knock", "Door open/slam", "Steps", "Chair moving", "Spoon/cup jingle", "Paper work", "Key jingle", "Keyboard typing", "Phone ring", "Applause", "Cough", and "Laugh".

3.2 Video signal annotation

Activity classification tier describes the considered interval in terms of the current activity in the room.

Movement annotation tier describes the current activity of each the participants.

Position of the participants is represented by 3D coordinates of each person (person head) in the room. Using 2D point markers from 5 cameras views together with camera calibration information the MSE estimator is applied to find the position of each person.

Focus-of-attention tier includes objects or subjects that each person is looking at. Note that we are actually annotating the individual gaze target, so the group focus-of-attention is included implicitly in the annotation. The gaze information of each person can be useful for extracting psychological aspects of the presentation: the level of interest of the lecture to the listeners, the degree of perception of the materials, etc. A minimum duration of 2 sec is considered for annotation.

Hand gestures include 2 categories: raising a hand to catch attention and pointing gesture. Raising a hand usually precede to the question

Head gestures also include 2 categories: agreement signal that usually corresponds to the movement of the head along vertical axis and disagreement signal that corresponds to the movement of the head along horizontal axis.

3.2.1 Spatial Role Labelling

Recently, a spatial role labeling (SRL) task has been introduced [6] as a language-independent annotation scheme for spatial expressions existing in a sentence. One of the most outstanding characteristics of that annotation language is that the same paradigm can be applied also to the image and video modality. The same representation model for extraction from video data enables the combination of multimodal features for better recognition and disambiguation in each modality [6].

The framework consists of a set of spatial roles based on the *t*heory of holistic spatial semantics (HSS) with the intent of covering all aspects of spatial concepts, and spatial relations. The semantic spatial components in HSS theory are trajector, landmark, motion, spatial indicator etc [13]. *Trajector* is the entity (object, person or event) whose location or motion is of relevance. *Landmark* is the reference entity in relation to which the location or motion of the trajector is determined. *Motion* indicates the perceived movement. In order to explain the type of the spatial relations between trajector and the landmark, spatial indicators are used. They can be divided into region, direction and distance spatial indicators

Table 3. The defined set of relevant trajectors, landmarks, motion and spatial indicators.

Trajector	Landmark	Motion	Spatial indicator		
			Region	Direction	Distance
person	*person*	*typing*	*in*	*Left*	*far*
laptop	*table*	*moving*	*into*	*Right*	*near*
paper	*coffee table*	*entering*	*on*	*Front*	*N meters*
cup	*blackboard*	*leave*	*at*	*in front of*	
phone	*laptop*	*sit down*	*along*	*Above*	
	window	*stand up*	*around*	*behind*	
	door	*Put*	*to*	*Above*	
	chair	*point*	*between*	*Below*	
	everybody	*greet*	*by*	*Down*	
	room	*open*	*inside*	*towards*	
	cup	*drink*			
		write			

Originally, this annotation scheme has been proposed for tagging of the natural language with spatial roles. In the case of video recordings there are no sentences to be tagged. This way it is necessary to define of a set of *spatial events* that will be tagged using the presented annotation scheme. Since we are interested in spatial information in the database, the *spatial events* could be associated with motion. We define a set of motions, trajectors, landmarks and spatial indicators (Table 3) that are relevant in our meeting-rooms scenario. The motions are always related to some person or object (trajector). The motion is considered with a reference to other object or person (landmark). The type of this relation is described by spatial indicator.

Consider a video fragment when PersonX moves to the blackboard to make a presentation. In this case the corresponding time interval will be annotated as following:

<Trajector > PersonX </ Trajector >
<Landmark > Blackboard </Landmark >
< Motion > move </ Motion >
< Spatial indicator > to </ Spatial indicator >

3.3 Audio-visual signal annotation

3.3.1 Emotion annotation using EMOTION ML

For annotation of emotions of the participants in the meeting we use the specification of Emotion Markup Language 1.0 [8] that aims to strike a balance between practical applicability and scientific well-foundedness. EmotionML markup must refer to one or more vocabularies to be used for representing emotion-related states. Due to the lack of agreement in the community, the EmotionML specification does not preview a single default set which should apply. Instead, the user must explicitly state the value set used. Taking into account relatively low richness of technical presentation content in terms of emotional states, four categories have been defined as shown in Table 2. The degree of prominence is evaluated along the following scale: 1 – slightly prominent; 2 – moderate prominence, 3 – highly prominent. The observation modality (audio, video and joint audio/video) is explicitly indicated in the emotion annotation.

3.3.2 Links between different layers of annotation

In the annotations the ambiguity in one modality is resolved by means of providing explicit links between audio and video tiers of annotation. As we already mentioned, the ambiguity in audio modality is represented by spatial and personal deixis or the missing answers to the questions (Table 1). In the case when deictic expressions have the corresponding video counterpart (pointing gesture, head gesture, focus of attention), an explicit link is provided between the two tiers of annotation.

3.4 Text annotation

3.4.1 Named entities

The named entity annotation task consists of three subtasks: *entity names*, *temporal expressions*, and *number expressions* [10] annotation. The *entity names* include “unique identifiers” of persons, locations, organizations. *Times* include dates, times, and durations. Finally, quantities include money amounts, measures, percents and cardinal numbers. Taking into account the specificity of the recorded material (scientific presentations), the new category “technical term” is also included that corresponds to the terms specific to the current topic of the presentation.

3.4.2 Topic annotation

In topic segmentation we follow the coding instructions for topic segmentation from the AMI meeting corpus [11]. The topic descriptions are divided into three categories:

1) Top-level topics. The content largely reflects the meeting structures.
2) Sub-topics. Divide top-level topics into more specific topics.
3) Functional descriptions. Generally refer to the very process and flow of the meeting, or are simply irrelevant. They are be further classified into: Opening, Closing, Chitchat (e.g. during coffee break).

4. DELIVERY

The final product is 5 hours of multimodal data together with its audio-visual annotation, that will be distributed by ELRA.

4.1 Assessing the quality of annotations

In general, reliability can be defined as a ”complex property of a series of observations or of the measuring process that makes it possible to obtain similar results if the measurement is repeated”. Since the degree of agreement between the annotators annotating the same recording can be considered as an indicator for reliability of annotation, inter-annotator agreement was computed for each tier separately. It is calculated as a pair-wise coincidence between two different annotations. Its calculation has been performed taking into consideration 6.6% of total amount of data annotated in the database. An average annotation agreement score 0.928 was obtained for the following four annotation tiers: “Movement”, “Hand gestures”, “Head gestures”, and “Activity classification” (see Table 4). The selected four tiers contain information that allows objective comparison between different annotators.

Table 4. Inter-annotator agreement

Tier	*Kappa*
Movement	0.953
Activity classification	0.952
Hand gestures	0.915
Head gestures	0.895
Average	0.928

5. CONCLUSIONS

In this paper, a multilingual database designed to develop technologies for rich audio-visual scene description in meeting-room environments was presented. . The database contains five hours of rather spontaneously generated scientific presentations which are manually annotated using standard or previously reported annotation schemes. Part of that database includes the already existing CHIL audio-visual recordings, whose annotations have been extended and upgraded. A relevant objective in the new

recorded sessions was to include situations in which the semantic content can not be extracted from a single modality.

The annotation process is performed from 3 main perspectives: annotation of audio signals, video signals and textual information. Regarding audio signals, orthographic transcription of speech corresponding to each speaker, the identity of speakers, and also 12 non-speech acoustic events, which naturally occur in meeting-room environments, has been performed. Regarding video signals, we annotated: body movements, position of the participants, gaze, focus of attention, hand gestures, head gestures, activity classification, and spatial role (with Spatial Role Labelling). Regarding textual information, named entities and topics were annotated. The quality assessment of annotation process was carried out for several annotation tiers and the results show acceptable results.

6. ACKNOWLEDGMENTS

This work has been funded by the Metanet4u EU project. The authors want to thank Jonatan Piñol, Alex Lopez, Diego Lendoiro, Coralí Planellas i Llongarriu, Emília Garcia for their participation in the new recordings. We are also grateful to the CHIL partners from AIT, ITC, IBM, UKA groups for their contribution to CHIL evaluation package.

7. REFERENCES

[1] D. Mostefa, N. Moreau, K. Choukri, et al, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms", Language Resources and Evaluation, pp.389-407, 2007.

[2] A. Waibel, R. Stiefelhagen, *Computers in the Human Interaction Loop*, Springer, New York, USA, 2009.

[3] A. Metallinou, C-C. Lee, C. Busso, S. Carnicke, S. S. Narayanan, Shrikanth S. "The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation", in Proc. Multimodal corpora: advances in capturing, coding and analyzing multimodality, 2010.

[4] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. Andre and K. Karpouzis, "A multimodal corpus for gesture expressivity analysis", in Proc. Multimodal corpora: advances in capturing, coding and analyzing multimodality, 2010.

[5] N. Tan, G. Ferré, M. Tellier, E. Cela, M.-A. Morel, J.-C. Martin, P. Blache, "Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation", in Proc. Multimodal corpora: advances in capturing, coding and analyzing multimodality, 2010.

[6] P. Kordjamshidi, M. Van Otterlo, M.-F. Moens "Spatial role labeling: task definition and annotation scheme", in Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010

[7] http://www.w3.org/TR/emotionml/

[8] http://metanet4u.eu/

[9] N. Chinchor, E. Brown, L. Ferro, P. Robinson, "1999, Named Entity Recognition Task Definition v1.4". ftp://jaguar.ncsl.nist.gov/ace/phase1/ne99_taskdef_v1_4.pdf

[10] http://corpus.amiproject.org/

[11] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Acoustic Event Detection Based on Feature-Level Fusion of Audio and Video Modalities," EURASIP Journal on Advances in Signal Processing, vol. 2011, Article ID 485738, 2011.

[12] H. Sloetjes, P. Wittenburg, "Annotation by category - ELAN and ISO DCR". In: Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008), 2008