Evaluation of multimodal components within CHIL:

The evaluation packages and results

Djamel Mostefa, Marie-Neige Garcia, Khalid Choukri
Evaluations and Language resources distribution agency (ELDA)
55-57, rue Brillat Savarin 75013 Paris, France
{mostefa, garcia, choukri}@elda.org

Abstract

This article describes the first CHIL evaluation campaign in which 12 technologies were evaluated. The major outcomes of the first evaluation campaign are the so-called Evaluation Packages. An evaluation package is the full documentation (definition and description of the evaluation methodologies, protocols and metrics) alongside the data sets and software scoring tools, which an organisation needs in order to perform the evaluation of one or more systems for a given technology. These evaluation packages will be made available to the community through ELDA General Catalogue.

1. Introduction

The project CHIL[1] – "Computers in the Human Interaction Loop" is an Integrated Project (IP 506909) funded by the European Commission under its 6th Framework Program. The project started on January 1st, 2004 and has a planned duration of three years. CHIL aims to radically change the way we use computers. Rather than expecting a human to attend to technology, CHIL attempts to develop computer assistants that attend to human activities, interactions, and intentions. Instead of reacting only to explicit user requests, such assistants proactively provide services by observing the implicit human request or need, much like a personal butler would. To achieve this goal, machines must understand the human context and activities better. This requires machines to better perceive and understand all the human communication signals including speech, facial expressions, attention, emotion, gestures, and many more. The research consortium includes 15 leading research laboratories from 9 countries representing today's state of the art in multimodal and perceptual user interface technologies in European Union and the US.

The first year of the project concluded with the First CHIL Evaluation Campaign in January 2005, followed by the First CHIL Evaluation Workshop, which took place in Athens on 20th and 21st of January 2005.

This abstract introduces the so-called Evaluation Packages, which are a major outcome of this 1st evaluation campaign. An evaluation package is the full documentation (definition and description of the evaluation methodologies, protocols and metrics) alongside the data sets and software scoring tools, which an organisation needs in order to perform the evaluation of one or more systems for a given technology. An evaluation package can be conditioned so that it can be shipped on a media to an organisation for it to reproduce one of the technology evaluations, which were conducted during the First CHIL Evaluation Campaign. These evaluation packages will be made available to the community through ELDA General Catalogue.

2. Databases

2.1. ISL Seminar 2003 database

The database contains audio and video recordings of 7 seminars given at the Interactive Systems Laboratories (ISL) of the University of Karlsruhe over the period of October and December 2003. These seminars are given by lecturers of the University of Karlsruhe or by invited speakers on topics concerning technologies involved in the CHIL project, such as speech recognition, audio source localization, audio scene analysis, video scene analysis, person identification and tracking… The language is European English spoken by mostly non native speakers.

2.1.1. Strategy for data collection

Collected data are audiovisual recordings taken from seminars given at the ISL, University of Karlsruhe.

From October to December 2003 and from April to July 2004, 2 weekly seminars were held, where students and other members of the computer science faculty presented scientific topics in the fields of speech recognition and multimodal user interfaces. The presentations were given in English and lasted about half an hour each. During the talks, videos of the speaker and the audience from 4 fixed cameras, frontal close ups of the speaker, close talking and far-field microphone data of the speaker's voice and ambient sounds were recorded to form a database of realistic data for development and evaluation of CHIL technologies.

2.1.2. Recording setup and material

Figure 1 gives an overview of the recording setup. The room in which the recordings took place is the computer room of the laboratory. Other than serving as seminar room, is it also widely used by students and members of the laboratory for programming, occasional meetings, discussions or video conferences. Its dimensions are 5.90x7.10m; the ceiling height is 3m. There is one entrance in the north wall and two more doors in the south wall leading to other offices.

Hardware and sensors. A number of sensors were installed in the CHIL room. These include: 4 fixed SONY DFW-V500 640x480 color firewire cameras in the room corners, at about 2.70m height; one pan-tilt-zoomable Canon VC-C1 640x480 color analog camera at the far end

of the room, opposite to the presenter area, to capture close ups of the speaker. Audio recordings of the presenter in the seminars were recorded with a close-talking microphone. In addition, a 2x8 channel microphone array was used to record audio in the room. For this array two 8-channel audio cards were used. All channels belonging to one of the audio-cards were synchronized on the byte level. The two 8 channel cards, however, were not synchronized at the byte level.

2.1.3. Audio transcriptions

For a single audiovisual data element (a seminar), two transcriptions were produced. The first one is the *speaker transcription* which contains the speech utterances of all intervening speakers, including human generated noises accompanying speech. This is done by transcribing the close-talking microphone recording of the main speaker. The second one is the *environment transcription* which contains all noises not produced by the speaker(s). *Environment transcriptions* are realized on far-field recordings. All the environmental noises (human and non-human) and all what the speakers say are transcribed.

Both transcriptions were produced with Transcriber[2] and are in native XML format.

2.1.4. Video labeling

Video annotations were realized using an *ad hoc* tool provided by University of Karlsruhe. This tool allowed displaying 1 over 10 pictures in sequence, for the 4 cameras. On each displayed picture, the annotator had to click on the lecturer's head "centroid", *i.e*. the estimated centre of his/her head if his/her face was visible.

The 2D coordinates of the hit point within the camera plane were saved to the corresponding camera's label file for further interpolation among all cameras in order to compute the real "ground truth" location of the speaker within the room.

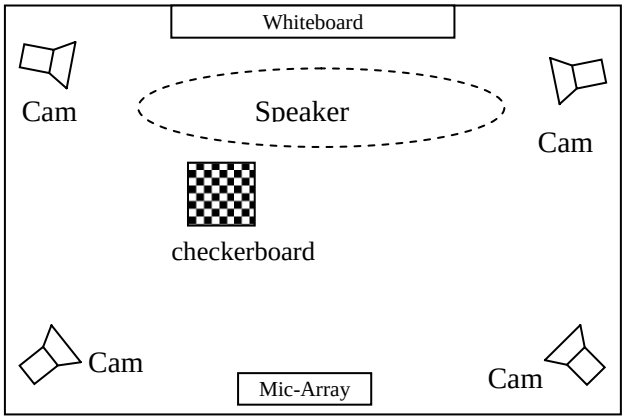

Figure 1 CHIL room setup at ISL

2.2. ISL Seminar 2004 database

This database is composed of 5 seminars recorded in November 2004. Like the 2003 seminars they are made of audio and video recordings of presentations given by researchers and students in the field of multimodality technologies. This database completes the ISL Seminar 2003 database and was used for the evaluation of audio and video technologies.

The same audio transcriptions and video labeling as described in 2.1.3 and 2.1.4 were carried on this database.

2.3. INRIA 2004 pointing database

The database has been produced by the French INRIA Rhônes-Alpes research laboratory.

The head pose database consists of 15 sets of images. Each set contains 2 series of 93 images of the same person at different poses. The first serie is used for training, the second is for testing. There are 15 people in the database, wearing glasses or not and having various skin color. The pose, or head orientation is determined by 2 angles which vary from -90 degrees to +90 degrees.

A detailed description of the database can be found in (Gourier, 2004).

This database was used for the Head pose estimation evaluation (see 3.1)

2.4. ISL Pointing 2003 database

The database has been produced by the University of Karlsruhe. Each video sequences shows a person performing pointing gestures. The recordings were done in 2 different places with a Mega-D stereo camera (Videre Design). The subjects are wearing a magnetic sensor (Flock of Birds) on the head that provides the rotation angles of the head. Furthermore, the pointing gestures as well as the positions of head and hands have been labeled manually. This database was used for evaluations of Head pose estimation and Hand tracking.

3. Technology components

During the first CHIL evaluation campaign, 12 different technologies were evaluated.

3.1. Vision technologies

Face detection. The goal in this evaluation process is to assess the quality and accuracy of the face/head detection and tracking techniques being used and developed within the context of CHIL. The evaluation was performed only taking into account the error between the centres of the detected faces with respect to the centres of the labeled faces in the ground truth data. To estimate the accuracy on size and extension of the estimated faces, we included in this evaluation a second metric that requires the estimation of the sizes of the detected and labeled faces. This new metric was used as a secondary evaluation metric.

Hand tracking. This task is about tracking the position of a person's left and right hand. The hand position is determined by the image coordinates (resp. 3D world coordinates) of the hand's centroid. Hand movement is an important feature for gesture recognition and human activity analysis.

Head pose estimation. The goal of this evaluation is to estimate a person's head pose. The head pose is determined by 3 angles: roll, pitch (also called tilt) and yaw (also called pan). The roll angle represents the person's head inclination with regard to the body, whereas

the pitch and the yaw stand respectively for the vertical and the horizontal inclination of the face. We did not evaluate the roll angle since the data does not contain much variation in roll. Knowing the head pose of a person provides important cues on his visual focus of attention, for example if the speaker is facing the audience.

Visual person tracking. The task in this evaluation is to find the position of a speaker giving a presentation in front of an audience in the CHIL room. The speaker position is determined by the 3D room coordinates of the speaker's head's centroïd. Three metrics were considered: *2D global mean error* (*e.g.* mean of Euclidian distance in millimeters between estimated position of the head centre, and the ground truth), *percentage of misses* (*e.g.* percentage of frames where no hypothesis is delivered, although a label exists) and *percentage of false positive (e.g.* percentage of frames where a position estimation is delivered, although no head label exists).

Visual Speaker Identification. Systems have to give an identity to each test segment. Test segments of 5, 10, 20 30, 60, 120, 300 seconds were considered. The training data consists of 5 frontal images of each individual. The correct classification metric was used. It computes the percentage of correctly identified segments.

3.2. Audio technologies

Acoustic scene analysis. This task used 18 semantic classes and 6 acoustic classes. The semantic classes are: breath, laughter, cough/throat, disagreement noise, conversation, generic mouth noise, speech, applause, beep, chair, door, footsteps, keys, music, papers, silence, typing, and electrical whirring. The acoustic classes are generic continuous tone, generic continuous sound without tone, generic single transient, generic regular repeated series of transients, generic irregular repeated series of transients, and other noise. The evaluation used only isolated instances of these sounds, and each sound had both an acoustic and a semantic tag (which could be "unknown"). The metrics, *mean per-class precision, mean per-class recall, and error rate*, were computed for two separate conditions: first, on the semantic tags; second, on the acoustic tags.

Acoustic Speaker Identification. Two tasks of speaker recognition are considered: speaker identification (SI) and speaker verification (SV). The first one consists in determining the identity of the speaker of a speech segment. In this task it is usually assumed that all the possible speakers are known. In the speaker verification task a speech signal and a proposed identity for the speaker are provided to the system, which has to determine if the proposed identity is correct or not. The correct classification rate metric was used to measure the results.

Acoustic Speaker Localization. The task involved in this evaluation corresponds in tracking the position of each speaker participating to a given lecture.

The task does not address situations in which two or more persons are speaking at the same time (competitive speakers).Moreover we dealt with both two-dimensional and three-dimensional source localization.

Evaluation was accomplished according to some reference transcriptions that were derived by video recording labeling as well as by manual transcriptions.

Close-talking microphone automatic speech recognition. The goal in this evaluation is to assess the quality and accuracy of speech recognition systems on the close-talking microphone recordings of the lecturer. The metric used was the word error rate.

Far-field microphone automatic speech recognition. Here speech recognition systems were evaluated on far-field recordings instead of close-talking microphone recordings. Far-field recordings include environmental noises as other speakers' speeches and are therefore more difficult to recognize. The word-error rate was used for comparison.

Speech activity detection. The goal is to segment an audio stream into speech and non speech segments. Far-field recordings were used as input. Systems' performances were measured with the *misclassification rate* (*e.g.* ratio between the duration of incorrect decisions and the total duration), *speech detection error rate* (*e.g.* ratio between the duration of incorrect decisions at speech segment and the duration of all speech segments), *non-speech detection error rate* (e.g. duration of incorrect decisions at non-speech segments / duration of non-speech segments) and the *average detection error rate* which is the average of the two previous metrics.

3.3. Content processing

Automatic summarization. The goal is to automatically extract summaries from oral transcriptions. A summary is extracted by selecting relevant chunks of words occurring in the oral transcription (*i.e.* units of summary). Different definitions of relevance and chunk were tested, as well as different sizes of summaries. The Translanguage English Database (TED) (Lamel, 1994) was used for this task.

4. Evaluation results and packages

For each technology, table 1 gives the result obtained by the best system during the official evaluation campaign.

For each evaluated component, an evaluation package is publicly available to the community through ELDA General Catalog[3]. It enables external players to benchmark their system and compare their results with those obtained during the official evaluation campaign. An evaluation package is the full documentation (definition and description of the evaluation methodologies, protocols and metrics) alongside the data sets and software scoring tools, which an organization needs in order to perform the evaluation of one or more systems for a given technology.

For each evaluated technology, a description of the protocol, metrics, tools, etc can be found in a CHIL public deliverable (Surcin, 2005).

Task	Metric	Result
Face Detection	*Percentage of correctly detected faces (%)*	97
Visual Person Tracking	*2D Global Mean Absolute Error (mm)*	130
Visual Speaker Identification	*Correct recognition rate (%)*	53.7
Head Pose Estimation	*Global Mean Absolute Error (%)*	12.4
Hand Tracking	*2D hand tracking error (pixel)*	15
Close-Talking microphone Automatic Speech Recognition	*Word error rate (%)*	23.6
Far-Field Automatic Speech Recognition	*Word error rate (%)*	51.9
Acoustic Person Tracking	*Localization rate (%)*	0.8
Acoustic Speaker Identification	*Misclassification rate (mismatch conditions) (%)*	12.9
Speech Activity Detection	*Average Detection Error Rate(%)*	12.2
Acoustic Scene Analysis	*Mean Per-Class Precision (%)*	26.7
Automatic Summarisation	*Recall-Oriented Understudy for Gisting Evaluation (%)*	0.24

Table 1 Best system results for each evaluated technology

5. Conclusion and future work

In this paper we presented the evaluation of 12 technologies that have been carried out during the first CHIL evaluation campaign in 2005. A major outcome of the campaign is the evaluation packages. For each evaluated technologies an evaluation package includes data, scoring tools, documentation, etc that enable an external site to evaluate its technology offline.
All these packages are made available through ELDA's catalog.

For the second CHIL evaluation campaign held in February 2006 the same kind of evaluation packages for audio, video and multimodal technologies (with similar content) is currently under production. Furthermore, other multimodal resources, not used in the official evaluations, and produced by CHIL partners will be also publicly available through ELDA's resources catalog.

6. References

Lamel L. et *al.* (1994). The Translanguage Englush Database (TED). In *Proceedings of International Conference on Speech and Language Processing (ICSLP) 1994.*

Gourier N. and Letessier J. (2004), The Pointing'04 Data Sets, *In Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, August 2004, Cambridge, UK*

Surcin S., Stiefelhagen R. and McDonough J. (2005), *D7.4 Evaluation Packages for the First CHIL Evaluation Campaign* http://chil.server.de/servlet/is/2712/

[1] http://chil.server.de
[2] http://trans.sourceforge.net
[3] http://catalog.elda.org