TM2: technical meetings

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

This resource is intended to upgrade the CHIL2007+ corpus with two newly recorded and annotated audiovisual technical meetings. The new recordings are in Spanish and Catalan. A relevant objective is to include situations in which the semantic content can not be extracted from only a single modality.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person:

Name: Climent Nadeu
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016438
Fax: +34 93 401 6447
e-mail: climent.nadeu@upc.edu

2.2. Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3. Copyright statement and information on IPR

The resource belongs to UPC. The resource is free, license-based, for research and commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive contains signal files, annotation files, and general documentation files, in a nested directory structure.

3.2. Encoding

Audio signals are in WAV format, at 44.1 KHz and 24 bits/sample.

Video signals are sequences of JPEG-compressed images, at a resolution of 752x582 pixels at 25 files per second.

Three different label file formats are used to annotate the corpus:

- XML format.
- ELAN XML [12] format.
- Text ASCII format.

Documentation is encoded in word/pdf/plain text.

3.3. Resource size (size of recorded data/MB occupied on disk)

The corpus contains about 175 GB of data.

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

The corpus is intended for doing research on video, audio and speech technologies, and in particular on the integration of their outputs for a multi-level based scene analysis. It was designed to upgrade and extend the already existing CHIL2007 corpus. While that existing corpus was in English, the new recorded data has been produced in two other languages, Catalan and Spanish. A relevant objective in the two new recorded sessions is to include situations in which the semantic content can not be extracted from only a single modality. Also, annotations have been largely extended with respect to the original CHIL2007 corpus.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The corpus consists of two technical meetings of about 30' each, one in Spanish and the other in Catalan. Four speakers talk rather spontaneously, and long speeches are usual. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, interaction among the attendees, coffee breaks, etc.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

The CHIL2007 corpus includes manual annotations from both audio and video modalities.

It contains a detailed multichannel verbatim orthographic transcription of the audio modality, which, besides speech transcription, includes speaker turns and identities, speech endpoints, vocalizations (e.g. <uh>, <uhm>, <Smack>, <B>), and acoustic events (from a set of 12 predefined events, like cough, laugh, door slam, chair moving,…). Also named entities and topics have been labeled from the ortographic transcriptions of speech.

Video annotations provide 3D multiperson head locations, movement, focus of attention, hand gestures, head gestures, and spatial role labeling (spatial relations).

From both modalities, annotations are included regarding activity classification, emotions, and links between different tiers.

The annotations were validated internally using inter-annotator agreement, by taking into consideration 6.6% of the total amount of annotated data.

4.4. Lexicon. Description of the lexicon (if applicable)

Not applicable

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

4 speakers per session, 6 speakers in total. The meeting participants were staff members and students. One speaker, who participated in both sessions was non-native in both languages.

4.6. Recording platform

4.6.1.Domain(s), environments,

Smart-room equipped with several cameras and microphones on the walls.

4.6.2. Recording platform

A set of audio sensors:

- 6 clusters of four-channel T-shaped omnidirectional microphone (24 microphones);
- 4 tabletop directional cardioid microphones;
- 4 close-talking directional wireless microphones.

A set of video sensors:

- four fixed cameras located at the room corners;
- one fixed, wide-angle panoramic camera located under the room ceiling.

For audio data capture, all microphones were connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers. The pre-amplifier outputs were sampled at 44.1 kHz and 24 bits per sample, and were recorded to a computer in WAV format via an RME Hammerfall HDSP9652 I/O card.

The cameras provide images of 752x582 pixels, and frame rate 25 fps. All video streams were saved as sequences of JPEG-compressed images.