FP6- 506811

AMI
AUGMENTED MULTI–PARTY INTERACTION
http://www.amiproject.org/

Integrated Project
Information Society Technologies

D3.3 ANNOTATED AMI HUB CORPUS

Due date: 30/06/2006 Submission date: 09/08/2006

Project start date:1/1/2004 Duration: 36 months

Lead Contractor: UEDIN Revision: 1

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D3.3 ANNOTATED AMI HUB CORPUS

Abstract:

In June 2006, the AMI project consortium released the AMI Meeting Corpus to the wider community. The corpus includes signals, transcription, and many different kinds of annotation. This document is a cover sheet for the data set that forms AMI deliverable D3.3. This cover sheet describes the corpus briefly and refers the interested reader to the deliverable itself, which can be accessed at http://corpus.amiproject.org.

Contents

1 Description of the delivered data set

The AMI Meeting Corpus is a multi-modal data set, or "corpus", consisting of 100 hours of meeting recordings. Around two-thirds of the data has been elicited using a scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The rest consists of naturally occurring meetings in a range of domains.

The meetings in the corpus have been recorded using a range of signals that are synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. During the meetings, the participants have instrumented pens and paper available to them. Pen outputs are available as part of the corpus but are not synchronized to the common timeline used by the other signals.

As well as the signals, the data set includes manually produced orthographic transcription of the language used during the meetings. This transcription is aligned at the word level with the common timeline. The corpus documentation includes the detailed instructions that were used by the transcribers to ensure consistency.

Finally, the data set includes annotations describing participant behaviour during the meetings at a wide range of levels. These include dialogue acts; topic segmentation; extractive and abstractive summaries; named entities; limited forms of head gesture, hand gesture, and gaze direction; movement around the room; emotional state; and where heads are located on the video frames. Not all 100 hours of meetings have been marked with all kinds of annotations, but the consortium has worked on data subsets of a size and shape they considered most useful. The linguistically motivated annotations have been applied most widely, and cover all of the scenario meetings. As for transcription, the documentation includes detailed instructions describing the annotations.

2 Obtaining access to the deliverable

This document is a cover sheet for the data set that forms AMI deliverable D3.3. The deliverable itself can be accessed at http://corpus.amiproject.org. Figure 1 shows a screenshot of the website's main page.

The website makes signal samples and documentation of the data set available anonymously to anyone with internet access. Users must register before accessing the corpus itself. This is so that we can have some idea who is using the data, and also so that we can be reassured that they have noticed the licensing conditions for data use. The registration process itself is simple, requiring only a name, email address, and confirmation that the license has been read before issuing a username and password for data access. After registration, users can access the signals from the corpus in a range of formats that differ in download size and serve the purposes of different types of users. Registration also gives access to the data annotations, including the orthographic transcription, and to a user forum. The bulk of the annotations have already been released on the website, but more are being released in stages, with the last public release of annotations produced using AMI funding scheduled for January 2007. All of our signals and annotations have been released under the terms of the Creative Commons Attribution NonCommercial ShareAlike 2.5 Licence. These terms state that if data users create new annotations and share them with others, they must release them publicly. This means that non-AMI funded annotations relating to the corpus may continue to be released beyond the project's end.

The website does not give access to full-size video signals, even though we have collected them and these are useful for video processing research. This is simply because these videos are too large for this access method. Instead, the website invites users who require them to contact us to arrange the shipment of firewire drives containing the data. The price for this service is set to cover production costs but not to make a profit.

Figure 1: Screenshot of the website that gives access to D3.3.