AMIDA

Augmented Multi-party Interaction with Distance Access

http://www.amidaproject.org/

Integrated Project IST–033812

Funded under 6th FWP (Sixth Framework Programme)

Action Line: IST-2005-2.5.7 Multimodal interfaces

Deliverable D3.3: Public Release of Annotated Remote Meetings: The AMIDA Meeting Corpus

Due date: 31/09/2008	**Submission date:** 07/10/2008
Project start date: 1/10/2006	**Duration:** 36 months
Lead Contractor: Mike Lincoln	**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)	

AMIDA

AUGMENTED MULTI-PARTY INTERACTION
WITH DISTANCE ACCESS

D3.3: Public Release of Annotated Remote Meetings: The AMIDA Meeting Corpus

Abstract:

In October 2008, the AMI project consortium released the AMIDA Meeting Corpus to the wider community. The corpus includes signals, transcription and some annotations. This document is a cover sheet for the data set that forms AMIDA deliverable D3.3. This cover sheet describes the corpus briefly with the corpus itself being publicly available at http://corpus.amidaproject.org/

1 Description of the delivered data set

The AMIDA Meeting Corpus is a multi-modal data set comprising 10 hours of recorded, transcribed and annotated meeting data, with a further 10 hours of signal-only data. The meeting data has a similar character to the scenario data in the AMI Meeting Corpus, but the AMIDA corpus contains meetings with one remote participant.

AMIDA meetings have 4 participants, each being assigned a role in the design of a new remote control. This scenario differs from the AMI corpus in that participants are asked to take over and finish the design project after another team has carried out the first two meetings. The participants can make use of all the material from these two previous meetings through a meeting browser as they prepare for and participate in meetings of their own. There are three four-person meetings of which two have a remote participant, plus one 2-person meeting where the participants are remote from each other. Communication for remote meetings is via video-conferencing.

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.

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 and is present for all meetings. The first data release also comprises a limited set of annotations which have passed the standard consortium 6-month embargo. A second release will be made in January 09 including all the annotations made to this point. Annotations that are present for at least some of the meetings include named entities, dialogue acts and topic segmentation, with more annotations to come including addressing, subjectivity, head and hand gestures.

2 Accessing 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.amidaproject.org/.

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. Some annotations have already been released on the website, but more are being released in stages, with the second public release scheduled for January 2009. 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.