

ASRU'2007, Kyoto, December 10, 2007

Recognition and Understanding of Meetings The AMI and AMIDA Projects

Steve Renals (1), Thomas Hain (2), and Hervé Boulard (3)

- (1) University of Edinburgh, Edinburgh, UK
- (2) University of Sheffield, Sheffield, UK
- (3) IDIAP Research Institute, Martigny, Switzerland



AMI and AMIDA EU projects



- 2 FP6 EU Integrated Projects, involving a dozen of EU institutions
- <http://www.amiproject.org>
- Analysis and modeling of (multimodal) human-human communication scenes
- **AMI: Augmented Multiparty Interaction**
 - Jan'04-Dec'06
 - Face-to-face meetings: analysis, modeling, retrieval (meeting browsers); offline processing
- **AMIDA: Augmented Multiparty Interaction with Distance Access**
 - AMI follow-on project, Oct'06-Sep'09
 - Online/offline facilitation of co-located and remote group collaboration
 - Any recording media, including commodity hardware
 - Applications: (1) Meeting support (engagement and floor control) and (2) archive access (content linking).

AMI-AMIDA Multi-Disciplinary Nature

Exploiting the nature of group interaction

- + speech and audio
- + non-verbal cues from video
- + attention focus, postures, expressions
- + **complementary multimodal cues**

multimodal nature



**social
psychology**



**computer vision
audio processing
a/v fusion**



multiparty nature

- + behavior constrained by group
- + role constrained by group
- + group size matters
- + **complementary multiparty cues**

AMI-AMIDA Multi-Sensor Meeting Room

Audio

- 4 close-talk microphones
- 4 lapel microphones
- 2 circular 8-microphone arrays

Video

- 4 close-view cameras
- 3 mid-view cameras

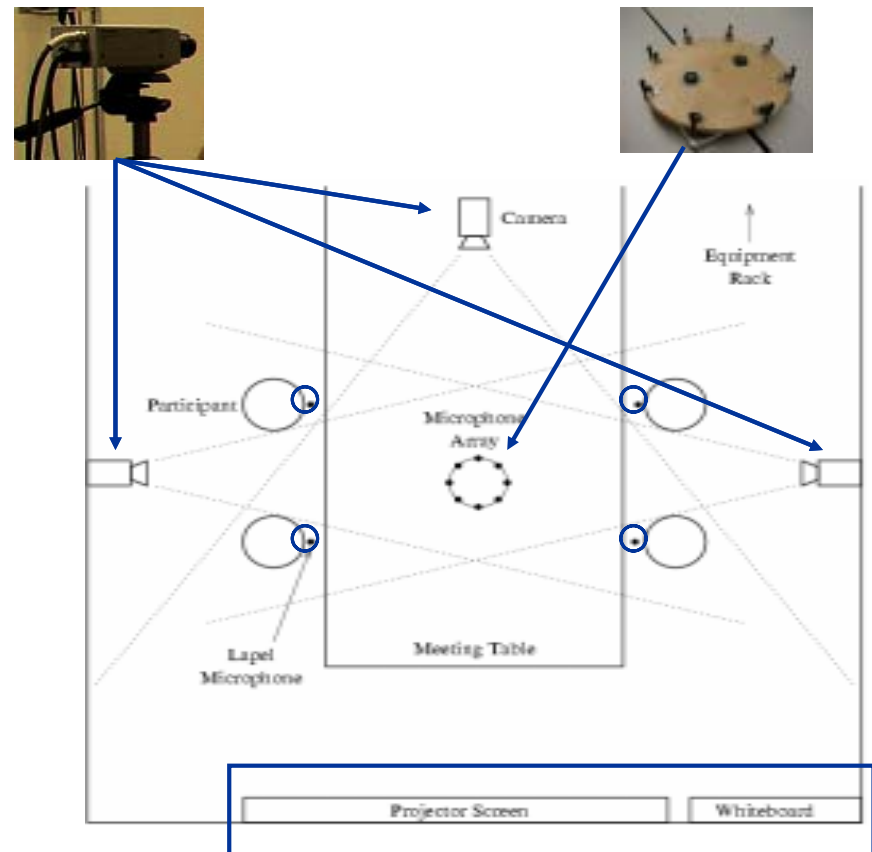
Slides

- projector screen

Handwriting

- whiteboard
- personnal notes

All time-stamped and synchronized



Corpus Design

- 70% role-plays of remote control design teams:
 - Task: design a remote control in and around four meetings
 - Roles: project manager, marketing expert, interface designer, industrial designer
 - Phases: functional design, conceptual design, detailed design
- 30% from a variety of genres (mostly real)
 - Find out where our methods and results generalize
 - Find out where they don't



IDIAP
(CH)



Edinburgh
(UK)

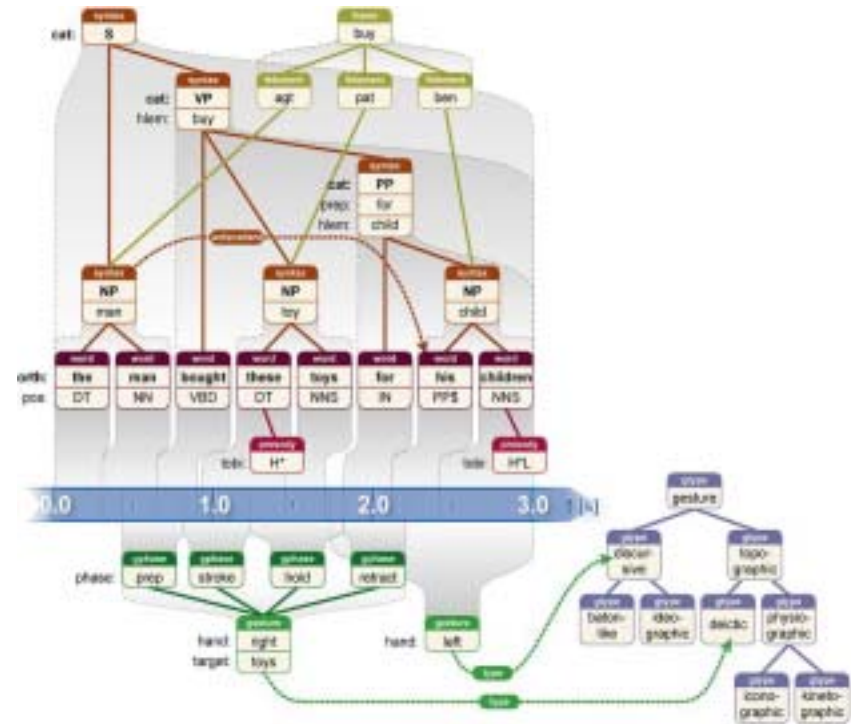


TNO
(NL)

Corpus Overview

- 100 hours of meeting data, manually annotated in terms of:
 - *Checked audio transcription*
 - *Topic segment./abs. summ*
 - *Named entities*
 - *Dialogue acts/ext. summ*
 - *Hand gestures*
 - *(limited) Head gestures*
 - *Location of person on video frame*
 - *(limited) gaze*
 - *Movement around room*
- ASR output (as well as other automatic processing outputs)
- Annotations carried out using NXT (the NITE XML Toolkit)
- Creative Commons Share-Alike licensing

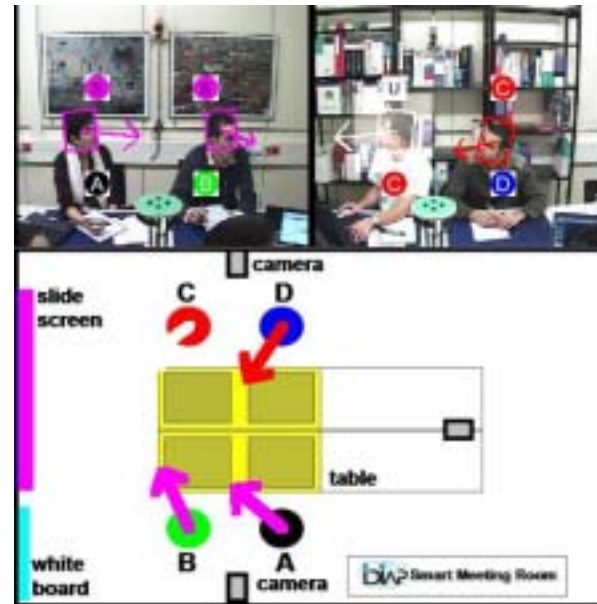
- Publicly available from <http://corpus.amiproject.org>
- DVD taster



Multiple layers/tiers,
hierarchical, support for
time aligned and general
content

Audio-Video Processing

- Active audio-visual shape tracking (multi-camera, multi-person tracking)
- Face and body tracking
- Face detection and recognition
 - Multi-view face detection (multiple tests on the AMI scenarios)
- Multi-person activity discovery
 - Joint estimation of the number of people, and their body and head location
 - Pose estimation, robust tracking
- Head activity analysis:
 - Focus of attention in meeting rooms
- Gesture and action recognition
- Handwriting and writer recognition:
 - Whiteboard handwriting recognition (extended database, better features)



Meeting Speech Recognition



- Basic properties
 - Conversational speech
 - Variable number of participants
 - Wide range of topics
- Challenges
 - Acoustic conditions: Far-field microphones, reverberation, ...
 - Multi-speaker: lively discussions lead to overlapped speech
 - (non-native speech) in most existing corpora

Meeting Speech Recognition

The Meeting Domain (1)

- Several corpora available (200 hours)
 - Mix of accents (US,UK,Intl.)
- Topics range from social to technical discussions on ASR

Meeting resource	Size (hours)	Avg Dur (sec)	Avg. Words/Seg
ICSI	60.1	2.1	7.3
NIST	20.9	2.3	7.2
ISL	7.8	2.4	8.8
AMI	84.1	3.3	10.1
VT	1.4	2.5	8.3
CHIL	(?)	1.8	5.6

Size and segment length statistic for several corpora

Meeting Speech Recognition

Meeting domain (2)

- Vocabulary differences

	Vocabulary source							
	ICSI		NIST		ISL		AMI	
ICSI	0.0	0.0	5.0	0.5	7.1	0.6	6.8	0.6
NIST	4.5	0.4	0.0	0.1	6.5	0.6	6.9	0.7
ISL	5.1	0.4	5.9	0.4	0.0	0.0	6.7	0.6
AMI	4.5	0.5	4.4	0.5	5.4	0.6	0.0	0.3
COMB	1.6	0.2	4.4	0.4	6.2	0.5	6.0	0.6

- Out of Vocabulary rates (OOV) **with padding to 50k words** from general Broadcast News data.
- Seemingly no need for specific vocabulary !

Meeting Speech Recognition

Front-end Processing



- Close-talking microphones
 - Speech activity detection is hard because of severe cross-talk.
 - AMI system uses MLP trained on 90 hours of speech and silence.
- Far-field microphones
 - Generic speech enhancement approach adopted
⇒ microphone placement information not required
 - Delay-sum beam forming
 - Wiener filtering
 - BIC based diarisation

Meeting Speech Recognition

The AMI System



- Essential system features
 - Multi-pass adapted system
 - System architecture independent of microphone source
 - CTS system adapted to the meeting domain
 - Phoneme posterior based features in front-end
 - Efficient meeting web-data collection
 - Discriminative training (MPE)

(for details see paper, Thomas Hain (NIST 2007))

Meeting Speech Recognition

AMI Performance



Description	Total	CMU	AMI	NIST	VT
Initial decode	37.4	47.7	29.3	33.8	38.4
Adapted	28.2	37.9	21.9	24.6	27.9
Best single output	25.4	34.5	20.4	21.1	25.3
Combined	24.9	33.9	19.8	20.9	24.7

Results on the NIST 2007 evaluation set, **close-talking mic**

Description	Total	Sub	Ins	Del
Initial decode	44.2	25.6	14.9	3.8
Adapted	38.9	18.5	16.8	3.5
Final	33.7	20.1	10.7	2.9
Manual segmentation	30.2	18.7	9.4	2.0

Results on the NIST 2007 evaluation set, **table-top mic**

Content Extraction

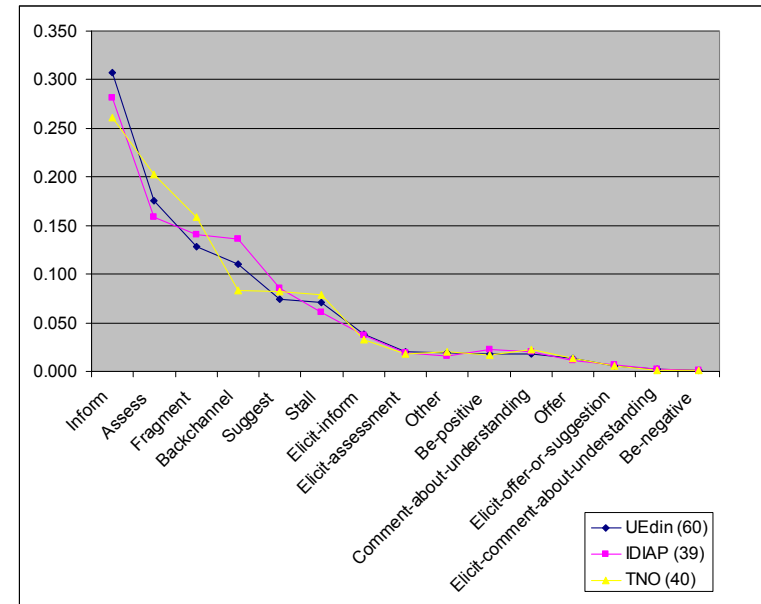


- Largely based on audio-video processing
- Emphasis on models and algorithms that combine modalities:
 - Extension of textual approaches to multimodal settings, involving the use of prosodic, video and contextual features.
- Main AMI-AMIDA contributions:
 - Dialogue act recognition
 - Topic segmentation/labeling
 - Dominance and influence
 - Summarization
 - Content-based automatic camera selection

Content Extraction

Dialogue Act Recognition

- **Dialogue acts: labels (15 in our case), defined for multi-party conversations:**
 - *Information exchange: giving and eliciting information*
 - *Possible actions: making or eliciting suggestions or offers*
 - *Commenting on the discussion*
 - *Social acts: expressing positive or negative feelings towards individuals or the group*
 - *Other: conveying an intention, but do not fit into the above categories*
 - *Backchannel, stall and fragment*
- **Can serve as elementary units, upon which further structuring or discourse processing may be based**



Our approach: switching dynamic Bayesian networks, modeling a set of features related to lexical content and prosody and incorporating a weighted interpolated factored language model

- *Possible to reach pretty low DA segmentation error rates*
- *But DA tagging and recognition remains a challenging tasks*

Content Extraction

Topic Segmentation










- Automatic inference of the sequential structure of the meeting
- Fundamental units (topics) are typically many minutes in duration
- Our approaches:
 1. Unsupervised (lexical cohesion): automatically infers topic boundaries as points where the statistics of text change significantly
 2. Supervised: allows to use additional features relating to prosody (e.g., pauses) and the structure of conversation (e.g., speaker overlap)
- Both topic segmentation and topic labeling are relatively robust to ASR errors.

Evaluation



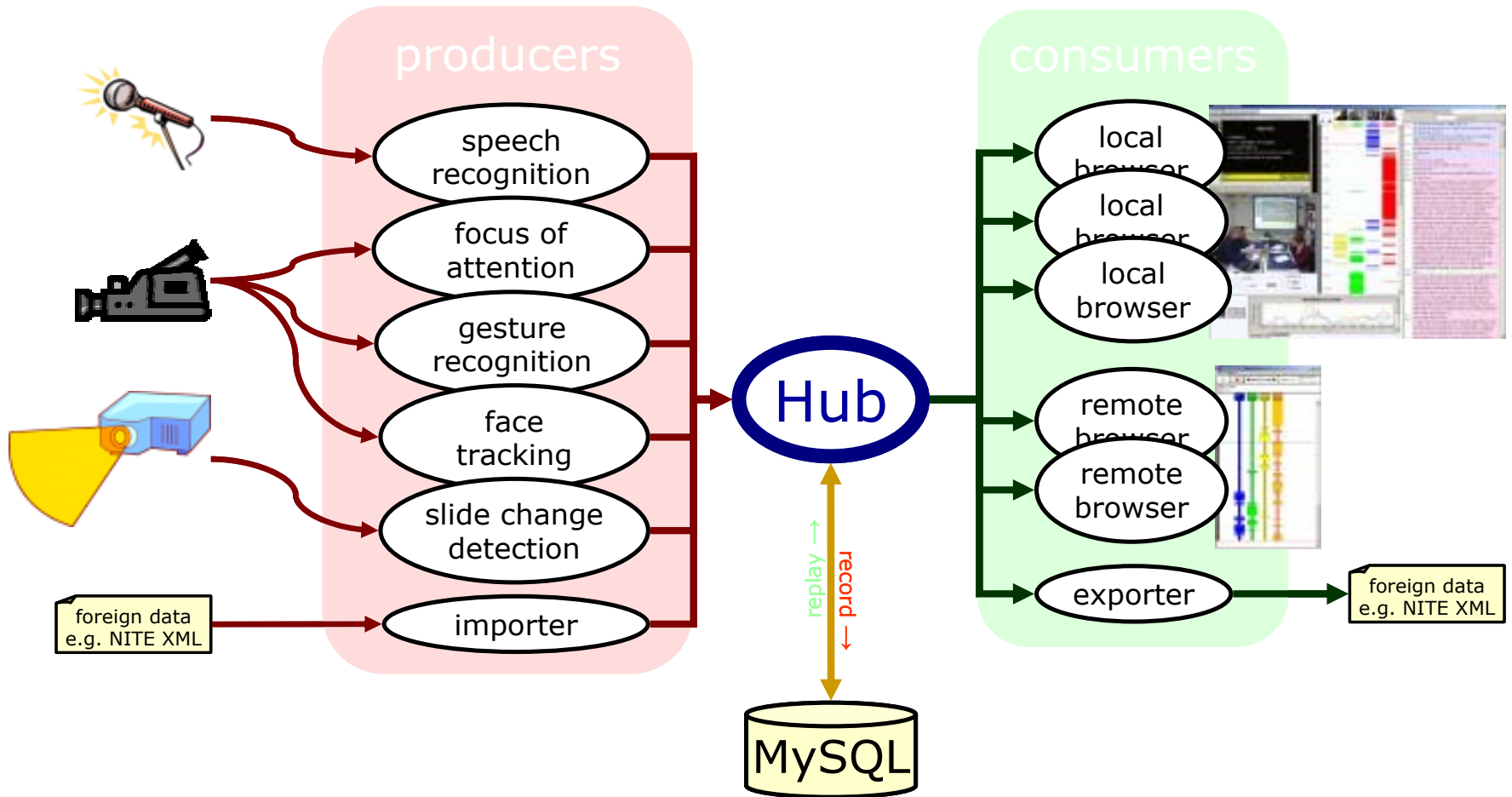
- **At the component technology level:**
 - Internal evaluations
 - NIST Meeting Recognition (RT) evaluations
 - CLEAR evaluations (VFOA and face detection)
 - AMI corpus, together with automatic ASR output, provided to the Cross Language Evaluation Forum (CLEF) for the 2007 evaluation on cross-lingual question answering

	Internal Evaluation (AMI/AMIDA)	External Evaluation	Contributing Data
ASR	✓	 NIST	✓
KWS	✓	 NIST	
SEG	✓	 NIST	✓
ID/LOC	✓	VACE (CLEAR)  	✓
FOA	✓	VACE-III  	✓
GAA	✓		

- **At the system level:** Summarization and topic segmentation: difficult to evaluate alone: may require extrinsic evaluation, in the context of scenarios and usability and utility testing.
 - One solution investigated so far: meeting Browser Evaluation Test (BET)

Conclusions

Current Status: The AMI "Hub"



Conclusions

From off-line (AMI) to online (AMIDA) processing

- **Technology and infrastructure:**
 - Commodity hardware (sensors) and lower bandwidth make speech recognition and computer vision much harder
 - Realtime processing
 - Realtime retrieval, integration, and distribution of all information available (metadata)
 - Integrating additional sensors/info sources (cell phones, IM)
 - Multiple interactive browsers (hence multiple SW clients)
- **Human interaction behaviours:**
 - Group size (less control)
 - Remote interactions (context awareness, presence)
 - Long-term collaboration patterns
- **Large amounts of unannotated data:**
 - AMI: 100 hours of annotated data
 - AMIDA: limited amount of training/adaptation data
- **Evaluation:**
 - More emphasis on system usability and utility, which is an issue by itself



Conclusions

From off-line (AMI) to online (AMIDA) processing



- **Engineering challenges**
 - Moving from ~30 times real time to ~1 times real time
 - Performance trade offs
 - Algorithmic delay (e.g. lattice rescoring)
 - Moving from specialised, expensive hardware to commodity, cheap hardware
- **Longer term**
 - Longitudinal processing: employing information from previous meetings
 - Investigating higher level processing/analysis (e.g., relationship between speaking style / speaker role / who they're talking to who), but this would require more participant overlap in future corpus collection
 - Social signal processing



Thank you!

Meeting Speech Recognition

Language modelling



LM component	Size	Weights (3-gram)
AMI (prelim!)	206K	0.04
Fisher	21M	0.24
Hub4 LM96	151M	0.04
ICSI	0.9M	0.08
ISL	119K	0.09
NIST	157K	0.07
Swbd/CallHome	3.4M	0.07
Webdata (meetings)	128M	0.16
Webdata(Fisher)	128M	0.10
Webdata(AMI)	138M	0.10

Language model components used for interpolation
and the associated weights