AMI
Augmented Multiparty Interaction

Steve Renals
Centre for Speech Technology Research
University of Edinburgh
AMI Research Vision

- Understanding human communication
  - Scene analysis
  - Unconstrained speech recognition
  - Model individuals and groups
  - Structure, index, summarize communication scenes
  - User interfaces
- Meetings provide a realistic, yet circumscribed arena to address these problems
Typical European project meeting

• Personnel: 2 person-months
  • Meetings: 15 people, for 15 hours.
  • Preparation: 15 people, for 5 hours.

• Travel budget: 6000 Euros
  • 15 airfares, at 250 € = 3750 Euro.
  • 15 nights in a hotel, at 100 € = 1500 Euro.
  • 15 dinners, 30 lunches, at 20€ = 900 Euro.
  • Total = 6150 Euro.

• Miscellaneous:
  • 2 days back-log of other work.
  • 30 nights away from family and friends.
  • 15 Friday evenings or Saturday mornings spent in transit in London airports.
Former FBI Special Agent Hosty’s contemporaneous handwritten notes from November 22, 1963 post-assassination interrogation of Lee Harvey Oswald.

F.P.C. in New York. Had applied for refu gees’ benefits from U.S. M.C. On 11/11, left for New York. On 11/12, worked at the consulate in Mexico. Had no electricity. Shuts down to go home because of confusion. 1:42 a.m. floor office, 3:45 a.m. floor, car storage. 4:05 a.m.
Questions

- Next week
  - “What happened at the review meeting?”

- Next month
  - “Did we discuss the new tracking algorithm with the people from Munich?”

- Next year
  - “What was the precise criticism of the XYZ work?”

- Today
  - “How could the others from Maryland have participated in an efficient way?”
AMI Objectives

- Technology to support human interaction in meetings
  - Meeting Browser
  - Remote Meeting Assistant
- Based on multimodal recordings from instrumented meeting rooms
  - Audio (multiple microphones - headmounted, lapel, tabletop arrays)
  - Video (closeup and room-view)
  - Slides (data projector capture)
  - Text (handwritten notes, whiteboard)
Instrumented meeting rooms

- AMI meeting rooms at IDIAP, TNO, UEDIN
- Standardized collection of 4 person meetings using:
  - 4 close-, 2 wide-view cameras
  - 4 headset, 8 array microphones
  - data projector capture
  - whiteboard capture
  - digital pen capture
  - extra site-dependent devices (eg second microphone array, lapel mics)
Instrumented meeting rooms
Component technologies

- Browsing meetings (online and archived) requires:
  - Models of group dynamics
  - Audio and video processing and recognition
  - Models to combine modalities
  - Content extraction
  - (As well as meeting user requirements and various software technologies)
- And lots of data... well annotated
Prototype Meeting Browser
AMI data collection

- Use cases for archive browsing and online assistants
- AMI scenario meetings: set of meetings on a common design project
- Data collection in the IDIAP, TNO, Edinburgh meeting rooms
- Hub corpus: 60% scenario meetings
- Spoke corpora: ICSI and M4 corpora; specific data for localisation and tracking
Annotation

- Annotation phenomena defined - cater for all the key research problems on hub corpus
- NITE XML format and toolkit to standardize annotations
- Annotations include:
  - speech transcription
  - dialogue acts
  - focus of attention
  - summarization
  - meeting actions
  - individual actions
Signal labeling
Dialogue Act Labeling

AMI Dialogue act coder

Transcription:
[Transcription content]

Adjacency pair:

Source:
Request Support

Target:
I think so

Inform

Edit Adjacency Pairs

Dialogue act:
Agent: B
DA type: <none>
DA text: my name is Francina

Edit Dialogue Acts

Acknowledgment

Inform

Requests

Suggests

Assessments

Social-Affective Acts

Unclassifiable

Set Comment

New DA

Delete
The main topic discussed was arrangements and objectives of an upcoming field trip to visit research partners (G1). A number of members reported their progress to date. If there are any tasks that some member can help others with, an overall description of the subproject, a multi-lingual speech recognition system for use by the cellular phone industry, along with consideration of some inconsistencies there. Specifically disk and speech issues.

Most important concerns are which combinations of features to use, and what combinations of languages and broad areas' corpora to use for the training.

The group will meet at the building at 8 am to go to the airport for a flight together.

Speaker m010 needs to discuss files that can be moved with speaker m007.

Further CMB would need to take a closer look at the subproject, and an estimate of how long the entire process should take.

At this meeting, they should discuss what they would ultimately do with the system.

People can discuss what m014 could do on the project to speed things up. Though creating the phrasebook seems possible, how can we integrate the phrasebook into the system? Speaker m011 is going to look into the machines that m017 had been running data on to find out what they are.

Rather than consider level of normalization as a further dimension for the project, whatever G01 finds the best will be used systematically.

Need to use multiple machines and SPERT boards to run processes as they take longer.

They will consider looking at articulatory features rather than just phonemes, though it wouldn't be perfect.

Speaker m007 has been preparing the French digit database.

Training and testing with varying noise.

Speaker m006 has installed updated software to everyone.

Working on live files from TIMT for relating neural nets.

Trying to figure out what the input to the system should be.

Speaker m012 has been testing the Italian database on a not tagged Spanish.

She has had problems with incorrect labels through.

Within the next year, the network is to be upgraded, and in a couple weeks, the group should have access to four new 36-processor workstations.

Multith has been copying some corpus stuff to a non-back up system, but not yet deleted originals.

Current plan is to use a superset of phones for the sub-project because from the various training languages.

HTTP errors currently takes 3 hours to a day, and the neural network takes 1-2 days.

It is not clear what combinations of dimensions, which features, should be run in the sub-project.

It is important to know that each speech processor has to be large and processor memory hungry.

To be aware in the fact that the cellular industry has an image of speech recognition in what they are after.

Must be careful if using a broad training set that is unfiltered and marked, because it would be unclear which is the reason for pronouncing.

Memory is of concern, because that product needs to run on a phone which has limited memory capacity.

GUI doesn't have a phrasebook properly prepared for them, so we're working with clusters, which may be good enough for digits, but not distinguishing words.
AMI PILOT RECORDINGS, SERIES 1

Please choose a meeting:

2004-08-02  AMI Meeting 001  With Ferret
(note: 3 seconds audio delay)

2004-08-02  AMI Meeting 002  With Ferret

2004-08-02  AMI Meeting 003  With Ferret

2004-08-02  AMI Meeting 004  With Ferret
Processing of audio-video data

- Defined according to core problems:
  - What did the participants say? (And how do they say it?)
  - What did they do? (physical actions)
  - Tracking each person’s location
  - The emotional state of each participant?
  - Tracking on what each participant is focusing
  - Who are the participants?
Structuring & content extraction

- Defined according to browser requirements
- Segmentation of multimodal streams
- Structuring by meeting events
- Identification of group activity
- Indexing and retrieval
- Summarization, and generation of textual and multimodal summaries
Multimedia presentation

- Presentation technologies for meeting data
  - JFerret browser (plugin integration framework)
  - Audio-only browser
  - New components for JFerret (eg browsing by slides)
  - Wireless presentation system
  - Virtual agents and environments for meeting playback
- Browser evaluation test
Current results

- Instrumented meeting room infrastructure
- Multimodal corpus of meeting recordings
- Meeting annotation schemes and tools
- Many component technologies: speech recognition, audio-visual tracking, summarization, ....
- Media file server
- JFerret meeting browser
- Open source software releases
  - NITE XML toolkit
  - TORCH machine learning toolkit