Recognition and Understanding of Meetings

Steve Renals
Centre for Speech Technology Research
University of Edinburgh

We spend a lot of time in meetings

Why study meetings?

- Natural communication scenes
  - Multistream - multiple asynchronous streams of data
  - Multimodal - words, prosody, gesture, attention
  - Multiparty - social roles, individual and group behaviours
- Meetings offer realistic, complex behaviours but in a circumscribed setting
- Applications based on meeting capture, analysis, recognition and interpretation

Why study meetings?

- Meetings offer a great arena for interdisciplinary research
  - signal processing
  - speech recognition
  - language and discourse processing
  - HCI
  - Social psychology

AMI

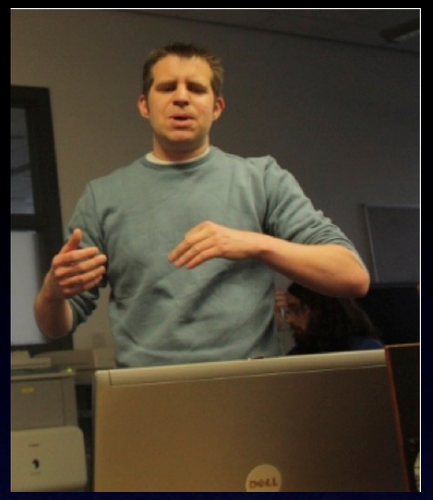
- Understanding human communication in meetings
- The AMI corpus
- Addressing challenges in interactive environments
  - multiparty, conversational distant speech recognition
  - meeting segmentation
  - meeting summarization
- Applications

AMI Corpus

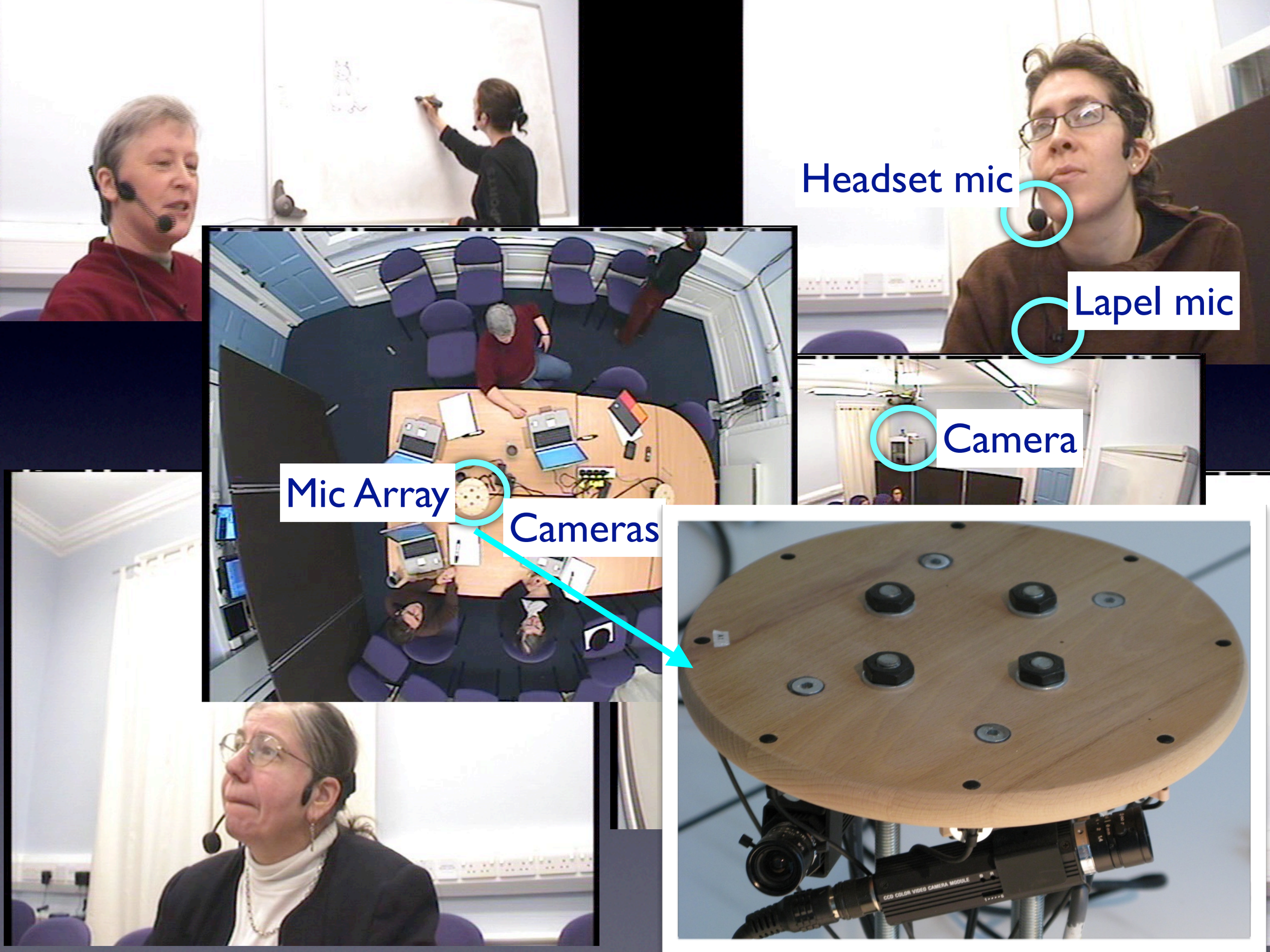
Recording multiparty interaction

- Two-party interaction
  - Switchboard
  - HCRC Map Task
- Multi-party interaction
  - ICSI Meetings
  - CMU ISL Meetings

AMI Corpus



- Multimodal multichannel meeting recordings
  - 70h ‘scenario-based’ meetings
  - 30h ‘non-scenario’ (real) meetings
  - 10h with remote participants (and using meeting browsers)


Headset mic
Lapel mic
Camera
Mic Array
Cameras

Scenario meetings?

- Scenario - team designing a remote control
- Each participant has a role (eg project manager)
- Roles stimulated by real-time email and web content
- Although scenario reduces overall realism
  - possible to define overall group outcome measures
  - controlled knowledge and motivation (no history)
  - can replicate the scenario (enable system-level evaluation)
- Recorded/annotated 30 replicates of the scenario

AMI corpus example

AMI Corpus

- Multimodal multichannel meeting recordings
  - 70h ‘scenario-based’ meetings
  - 30h ‘non-scenario’ (real) meetings
  - 10h with remote participants (and using meeting browsers)
- Manual annotations
  - linguistic: transcripts, topics, summaries, dialog acts, entities
  - multimodal: hand/head gestures, head pose, person location
- Automatic annotations: transcripts, topics, ...
- Creative Commons Attribution NonCommercial ShareAlike 2.5 License http://corpus.amiproject.org

Video labelling in NXT

Dialogue act labelling

AMI Dialogue act coder

File View Help

Transcription

IS1009a.sync.4:A: Okay . Request Support: <Everybody ready > ? Uh I think the first thing we do is introduce ourselves
IS1009a.sync.324:D: Inform: <Okay > .
IS1009a.sync.229:C: [vocalsound]
IS1009a.sync.326:D: Inform: <I think so > .
IS1009a.sync.328:D: Yeah , that's a good plan .
IS1009a.sync.6:A: and everybody's name and what your function is ? So maybe we start with you ?
IS1009a.sync.154:B: Okay . Yeah , Dialogue-act: <my name is Francina > . And I'm uh an user interface [disfmarker] my
esponsibility is user interface . And my role is to desig
IS1009a.sync.8:A: Mm-hmm .
IS1009a.sync.10:A: Mm-hmm . Ok
IS1009a.sync.330:D: And I'm the n
out how to sell them .
IS1009a.sync.12:A: Mm-hmm . An
IS1009a.sync.332:D: My name is E
IS1009a.sync.14:A: Okay .
IS1009a.sync.231:C: Yeah . Uh I'm
trial designer and my responsibilities are uh um um
sound]
technical-functional designs and specifications of user interface and dealing with user interface design .
IS1009a.sync.16:A: Very good . And as you already know I am Betty . I am the project manager for today . So why don't we look at the presentation [vocalsound] to see what we really are supposed to do .
IS1009a.sync.334:D: Okay .
IS1009a.sync.18:A: [vocalsound] Um .
IS1009a.sync.20:A: [vocalsound] Yes y opening , acquaintance , tool training . Well , the tools are , I think , we already [disfmarker] I guess the tool is really our [disfmarker] the computer , as far as I can see .
IS1009a.sync.336:D: Mm-hmm .
IS1009a.sync.21:A: Uh we get ins each of us will get instructions and we'll take it from there . Project plan , that falls under the same heading pretty much . Um , I don't think we have any great discussion at this point . Um .
IS1009a.sync.338:D: No .
IS1009a.sync.23:A: Here is what this thing should be . This thing we are gonna um uh design uh is a new remote control . Uh should be original [vocalsound] , trendy , and , of course , user friendly .
IS1009a.sync.340:D: Mm-hmm .
IS1009a.sync.342:D: [vocalsound]
IS1009a.sync.25:A: So maybe you wanna make some notes of that

- Acknowledgment
- Informs
- Requests
- Suggests
- Assessments
- Social-Affective Acts
- Unclassifiable

Edit Adjacency Pairs

Adjacency pair

Source... A: Everybody ready — Request Support
Type... POS
Target... D: I think so — Inform
Set Comment
New... Delete

Adjacency Pairs

IS1009a.adjacency-pairs.1:
-- Request Support
--- POS
-- Inform
=)--------
IS1009a.adjacency-pairs.2:
-- Request Support
--- POS
-- Inform
=)--------

Edit Dialogue Acts

Dialogue act

Agent: B
DA type: <none> — Type...
DA text: my name is Francina — Range...
Addressee: — All
Reflexivity:
Set Comment
New DA — Delete!

Nods

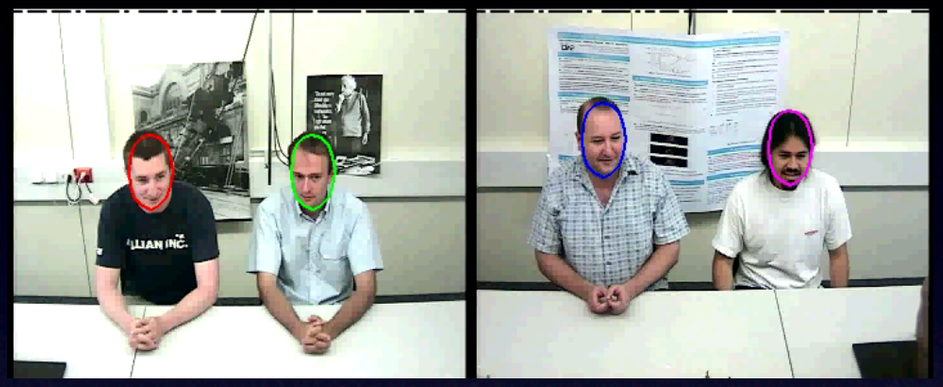
($d dact)($t da-type)($h head):($d>$t) & ($t@name="bck") & ($d # $h) & ($d@who= $h@who) & ($h@type="concord_signal")

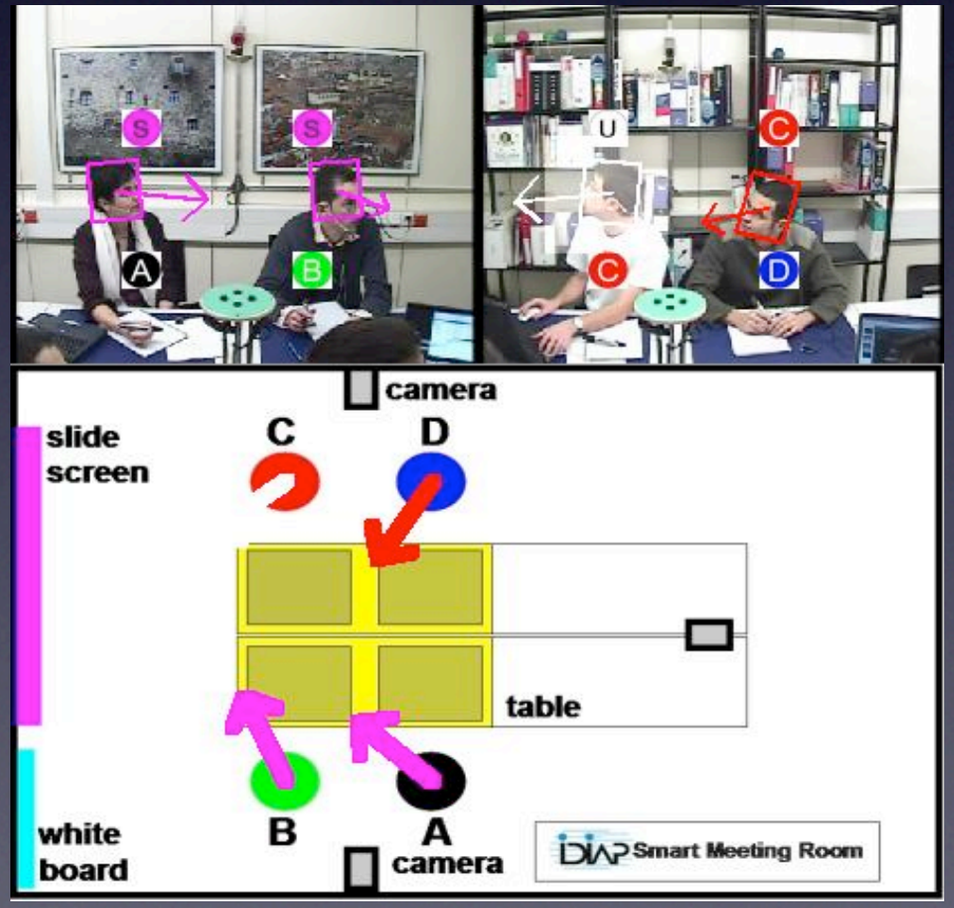
Gesturing while speaking

Recognition

Multimodal Recognition

- Speaker diarization
- Multi-camera tracking
- Activity discovery
- Head pose and visual focus of attention
- Multi-view face detection and recognition
- Gesture and action recognition





Speech Recognition

“ASR Complete” problem

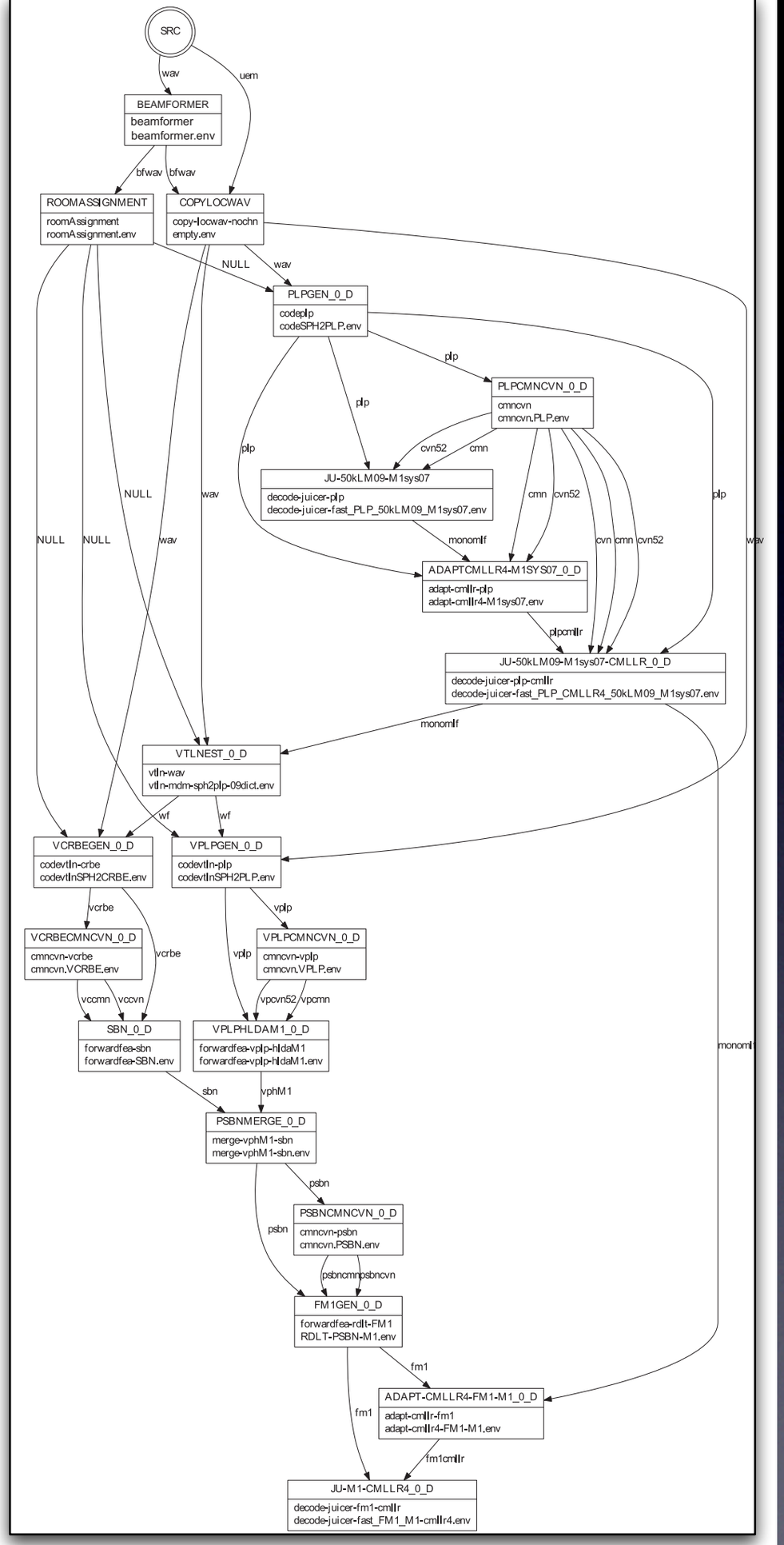
- Transcription of conversational speech
- Distant speech recognition with microphone arrays
- Speech separation, multiple acoustic channels
- Reverberation
- Overlap detection
- Utterance and speaker segmentation
- Disfluency detection

AMI-ASR

- Team from Sheffield, Idiap, Brno, and Edinburgh
- Acoustic preprocessing and enhancement depends on mic conditions
  - Individual headset mics (IHM)
  - Multiple distant mics / mic array (MDM)
- Multipass recognizer
  - HMM/GMM Acoustic model
  - n-gram language model
- No magic bullet for high accuracy ... more like an acronym shotgun

Basic system

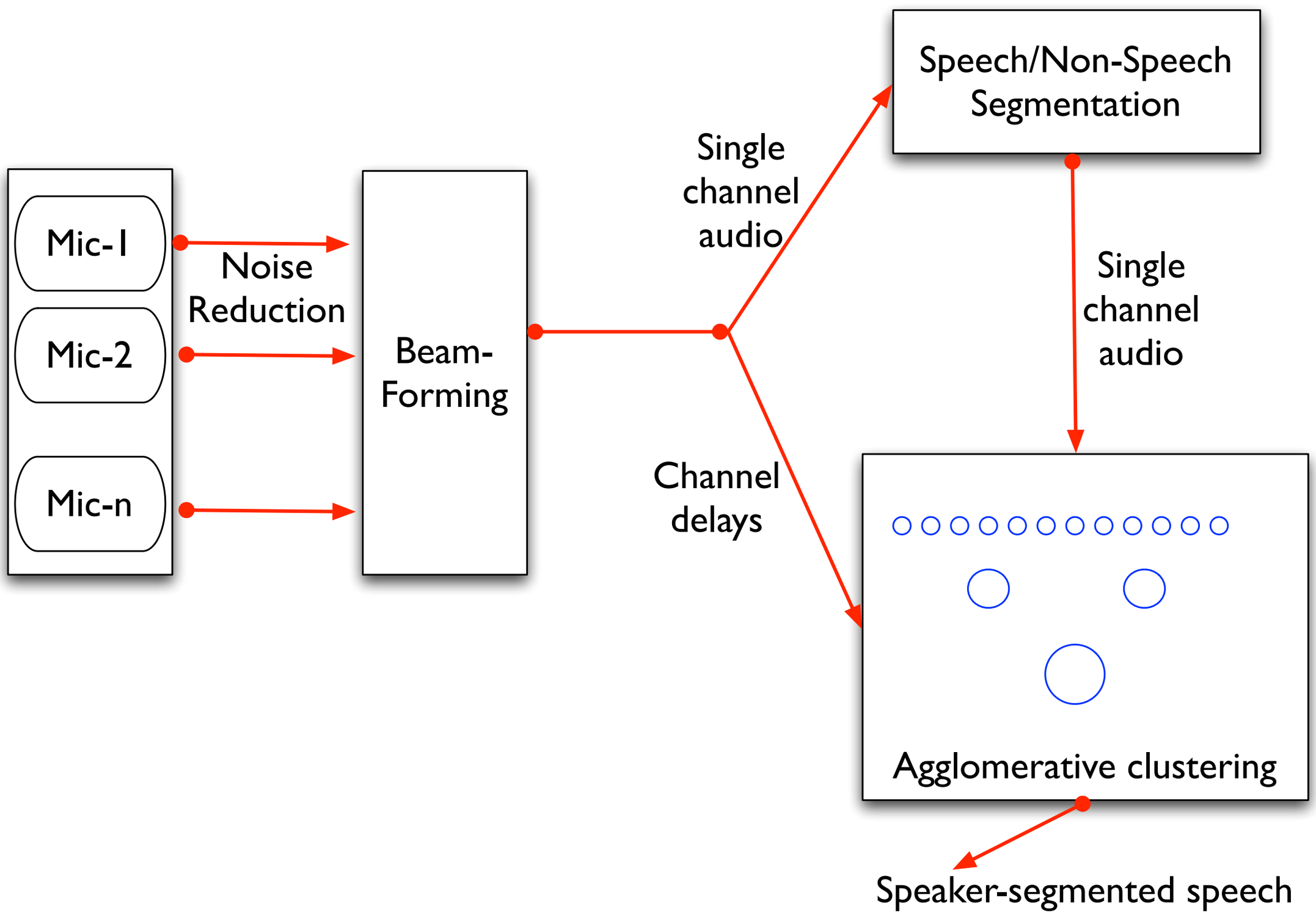
- Speech/non-speech segmentation
- PLP/MFCC features
- ML trained HMM/GMM system (122k 39D Gaussians)
- 50k vocabulary
- Trigram language model
- Weighted FST decoder



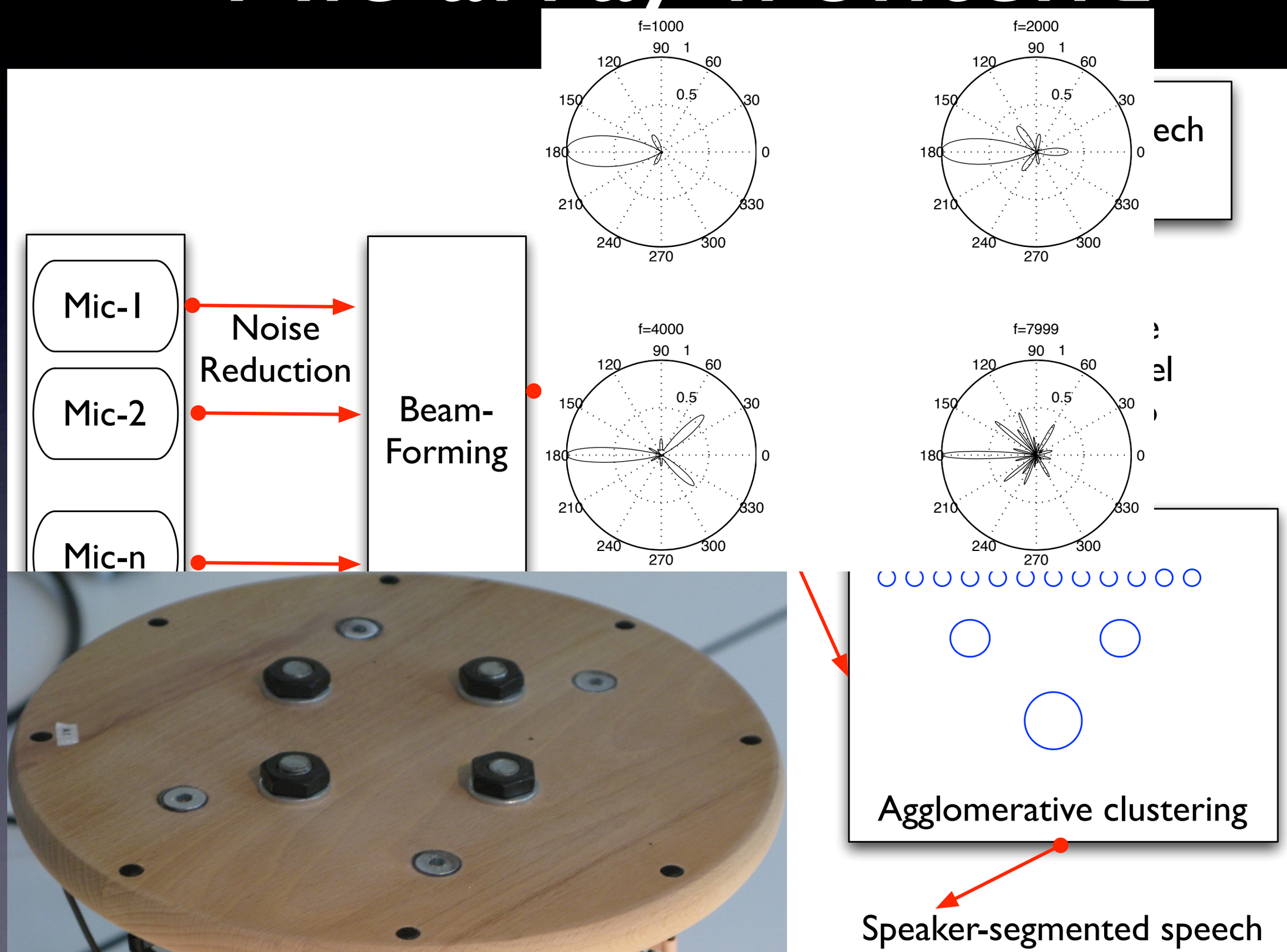
Additional components

- Microphone array front end
- Speaker / channel adaptation
  - Vocal tract length normalisation (VTLN)
  - Maximum likelihood linear regression (MLLR)
- Discriminative training
  - minimum Bayes risk (eg minimum phone error - MPE)
- Discriminative features and feature transforms
- Model combination

Mic array frontend



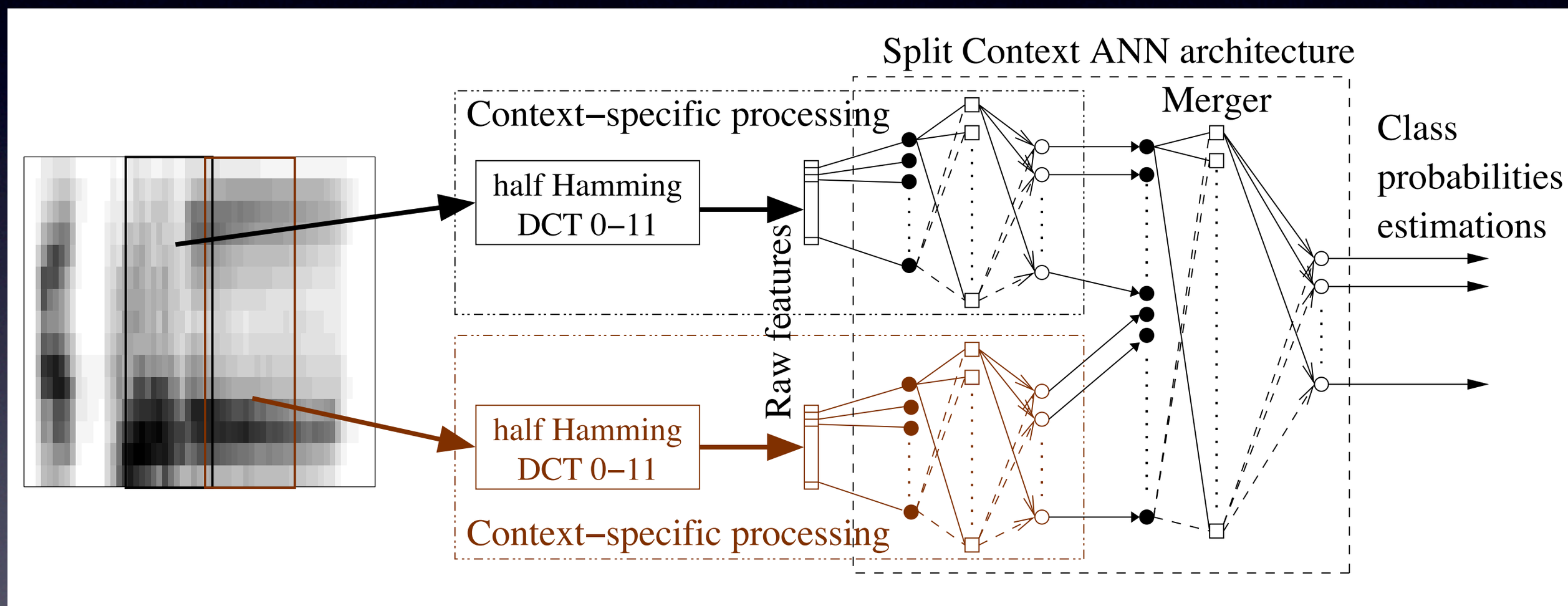
Mic array frontend



Discriminative features

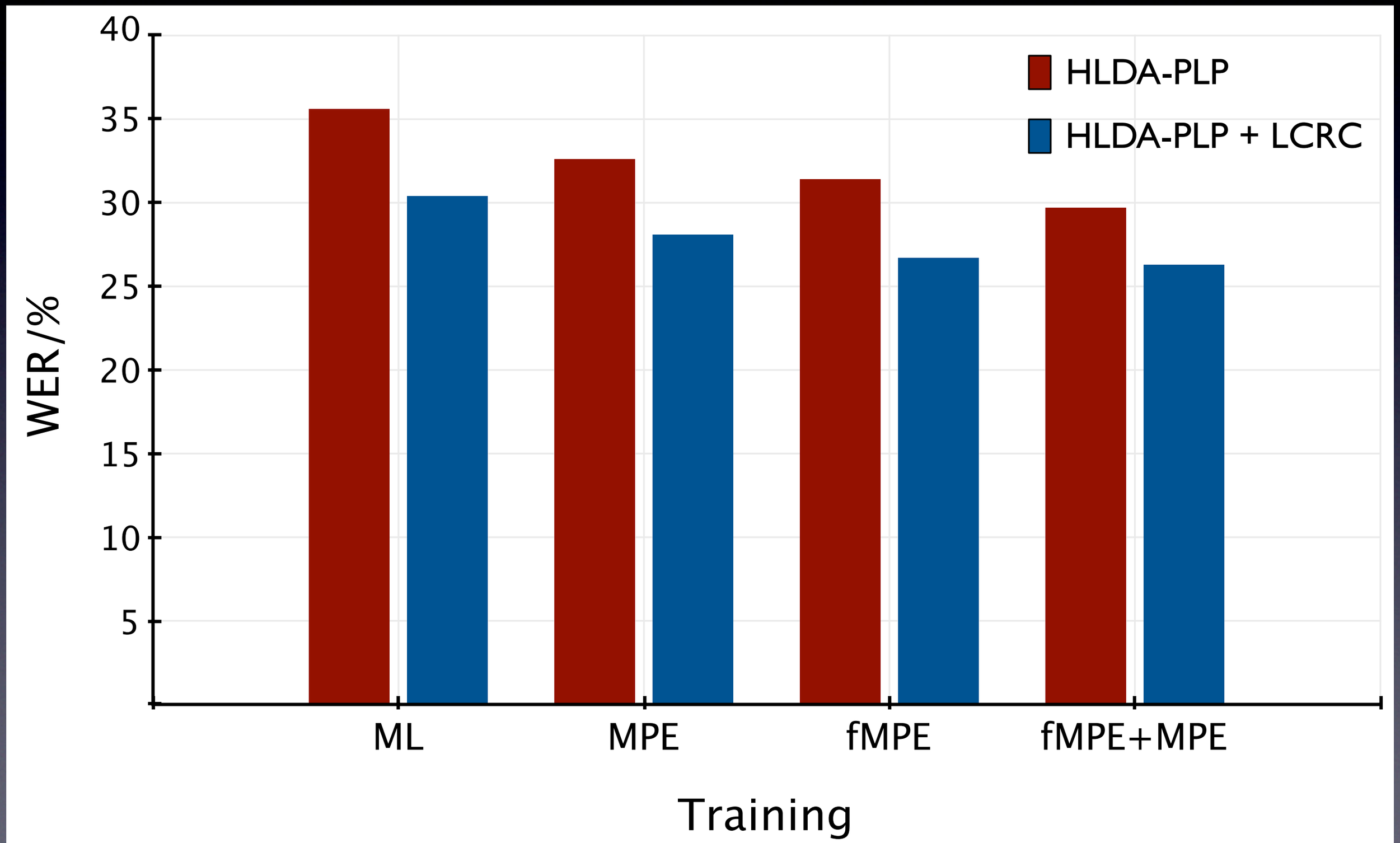
- Intensive amount of work on discriminative feature transforms (eg (H)LDA, fMPE)
- Posterior-based features from MLP phone classifiers
- Use as an additional feature stream
- Advantages
  - temporal context (±25 frames)
  - encode phone discrimination information
  - weakly correlated with PLP/MFCC features

LCRC features

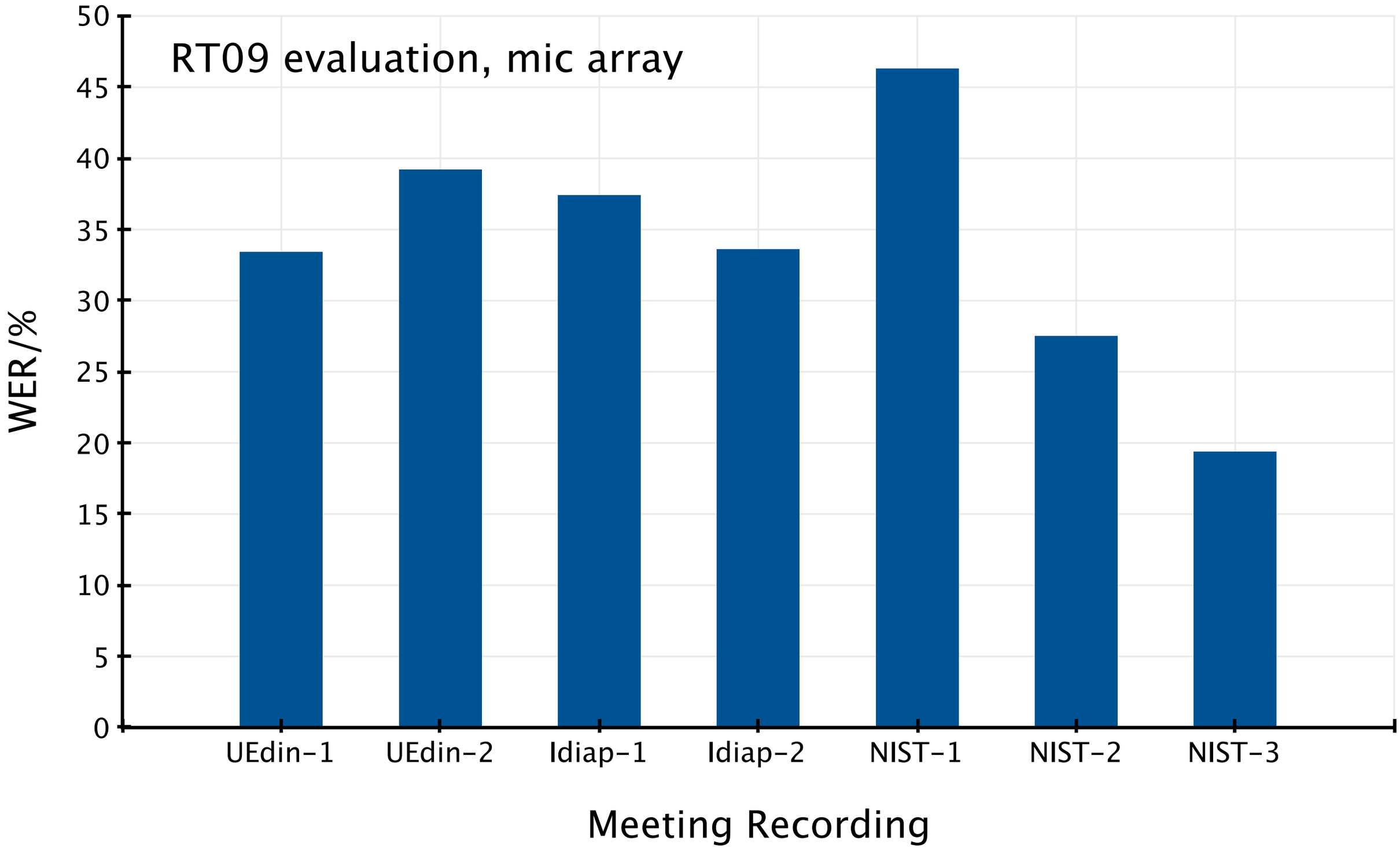


Karafiat, Grezl, Burget, Cernocky

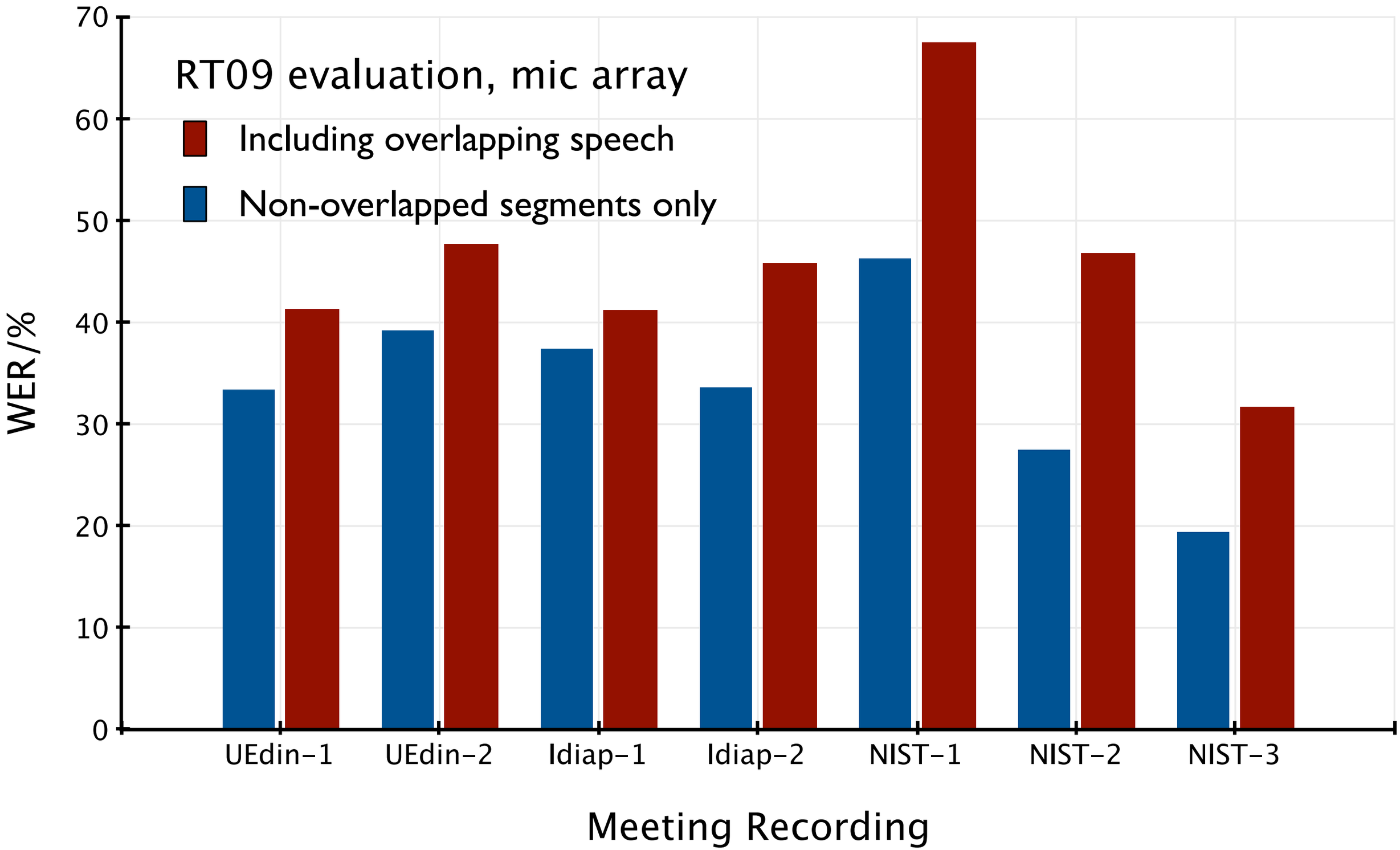
Results (RT07, IHM)



Results (RT09, MDM)



Results (RT09, MDM)



Meeting Interpretation

Meeting Segmentation

- Automatically segment meeting at different levels
  - dialogue acts
  - speaker
  - topic
  - meeting events
- Supervised and unsupervised methods
- Multimodal features: textual (ASR), prosodic, interaction, video

Meeting events

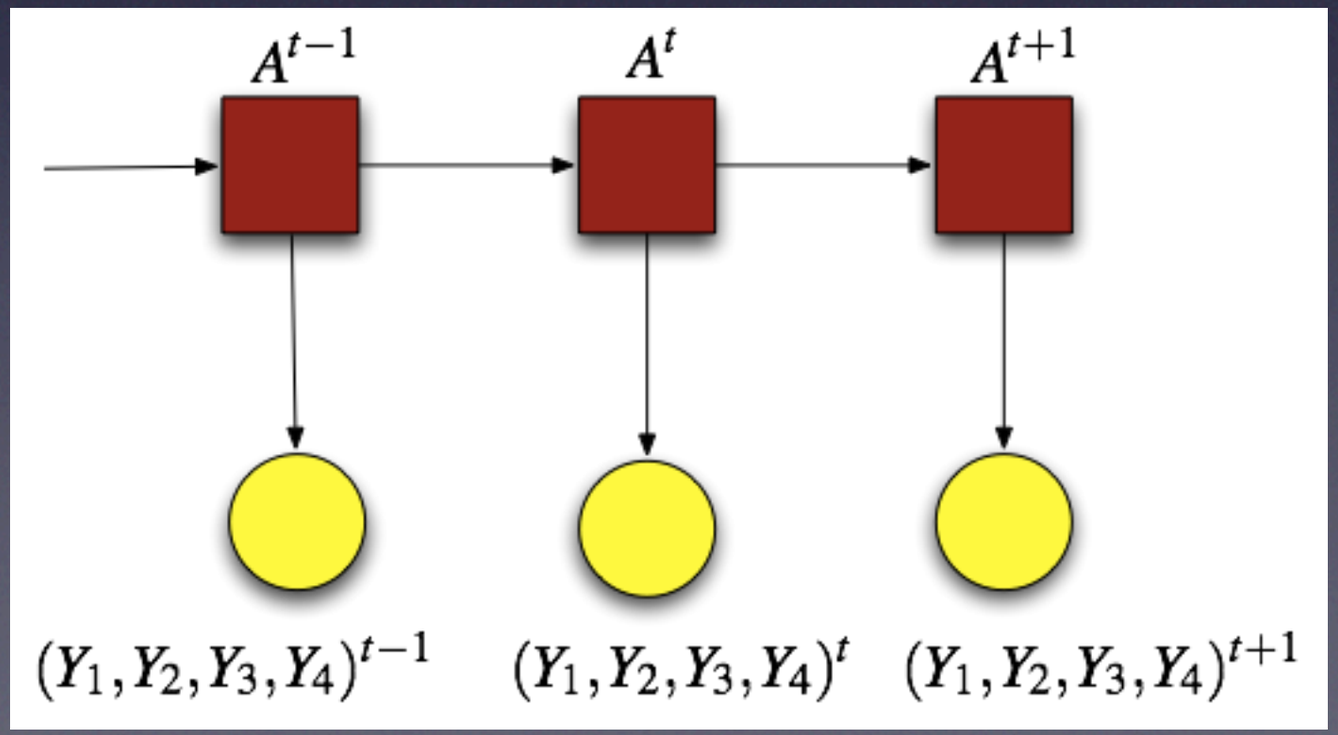
- Combine feature streams (speech, video, handwriting) to predict events in meetings
- Pilot study detection of meeting actions from a set of recorded meetings (M4 project)
  - Monologue
  - Discussion
  - Presentation
  - Speaking at whiteboard
  - Notetaking

Multimodal features

- Information is spread across individuals, modalities, sensor outputs
- Four sets of features:
  - Prosody (F0, rate of speech, energy)
  - Speaker turn features (speech activity in each of 6 locations, over 3 time periods)
  - Lexical features (trigram language models for different meeting phases)
  - Visual features: motion intensity and direction of skin-like blobs

Baseline model

- Define an HMM for each meeting action
- Each hidden variable generates the entire set of features (early integration)
- Gaussian mixture model pdf



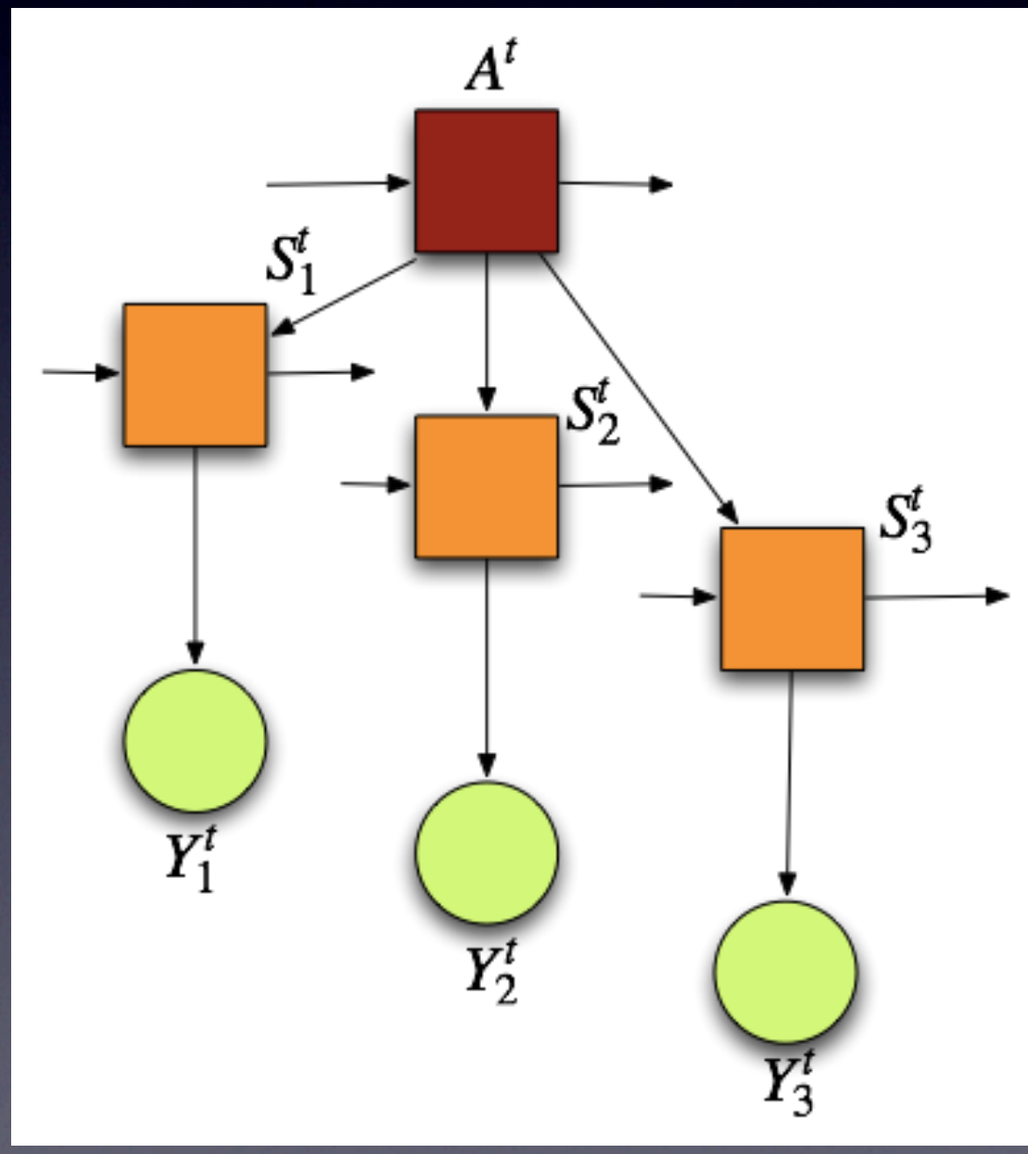
Baseline results

- Measure using Action Error Rate (based on sequence of correct actions

Spkr Turn Feats	55.1%
Lexical Feats	48.7%
Prosodic Feats	59.6%
Acoustic Comb	44.2%
Visual Feats	59%
MM Comb	43.6%

Multistream dynamic Bayesian network (DBN)

- Meeting actions decomposed as sequences of hidden subactions
- Multiple streams of subactions
- Richer hidden structure, distributed state representation
- Feature streams processed independently and asynchronously



Multistream DBN results

- Results on same task using 3-stream DBN, with 5 subactions per stream
- Counter enhancement is a way to model action duration

HMM	43.6
Multistream	13.5
Multistream + duration model	12.2

- We have used a similar model (more sophisticated LM) for dialogue act segmentation

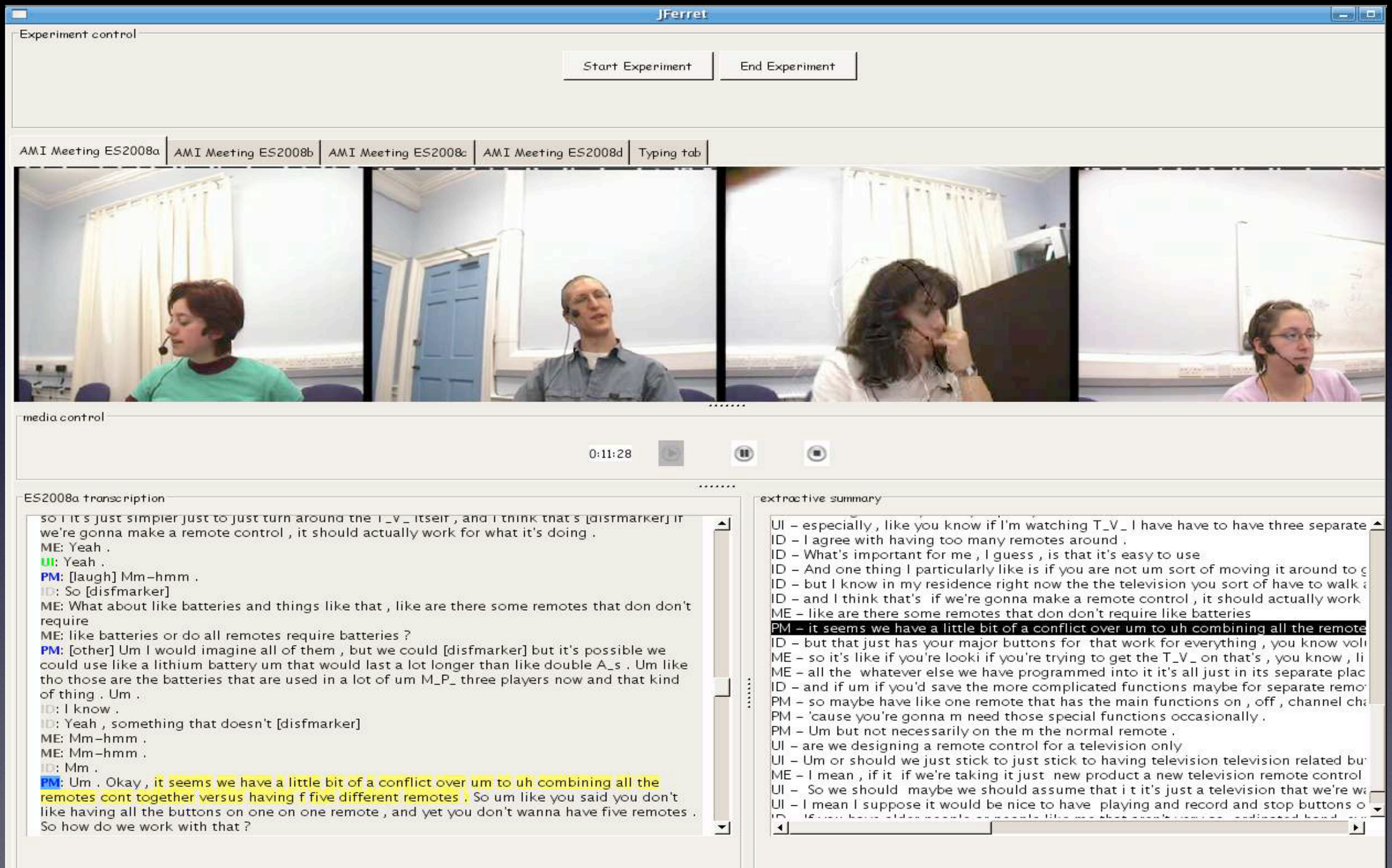
Summarisation

- Motivations
  - shield users from 30% WER transcripts!
  - decision audit, and other meeting review applications
  - (real-time) summarisation for collaborative environments
- Extractive summarisation
  - based on usual IR measures
  - also speaker-based measures
  - prosodic features
  - unit of summary - dialogue acts; speech ‘spurts’
  - use of multiple ASR hypotheses (word graphs)
  - sentence compression, disfluency removal

Evaluating summarisation

- Low correlation between ROUGE and human judgements
- Subjective decision audit evaluation
  - Comparing summarisation-based browsers to find why a decision was made
  - Objective and subjective evaluation measures
  - Compared browsers based on:
    - ASR vs hand transcripts
    - Keywords vs extractive vs abstractive vs hand summaries
  - 50 subjects

Summary-based browser

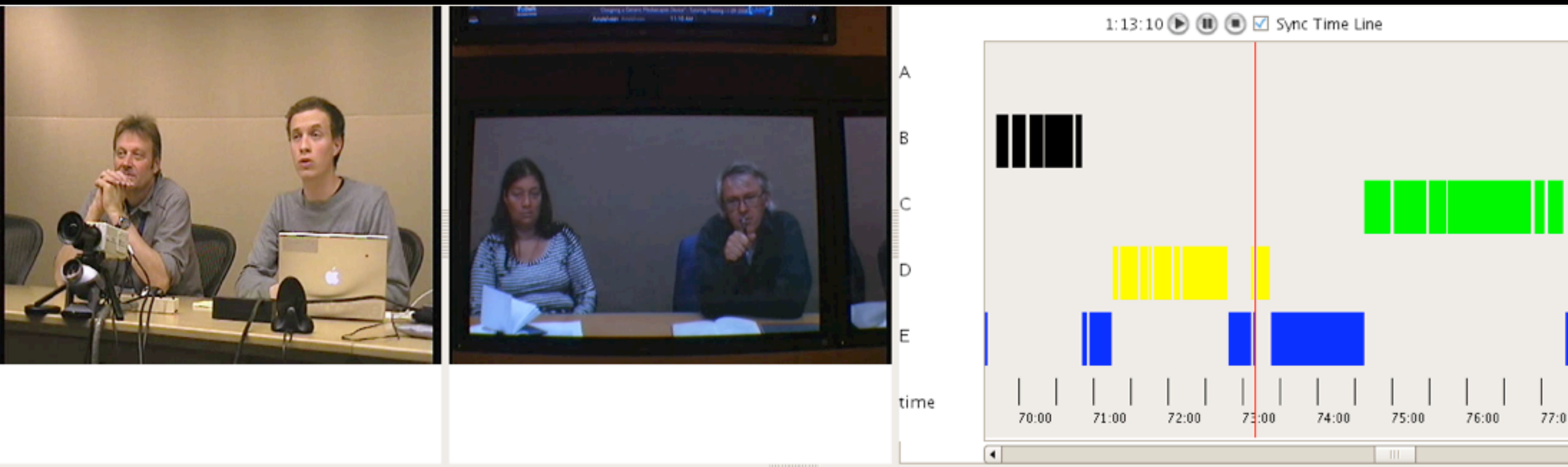


Decision audit evaluation

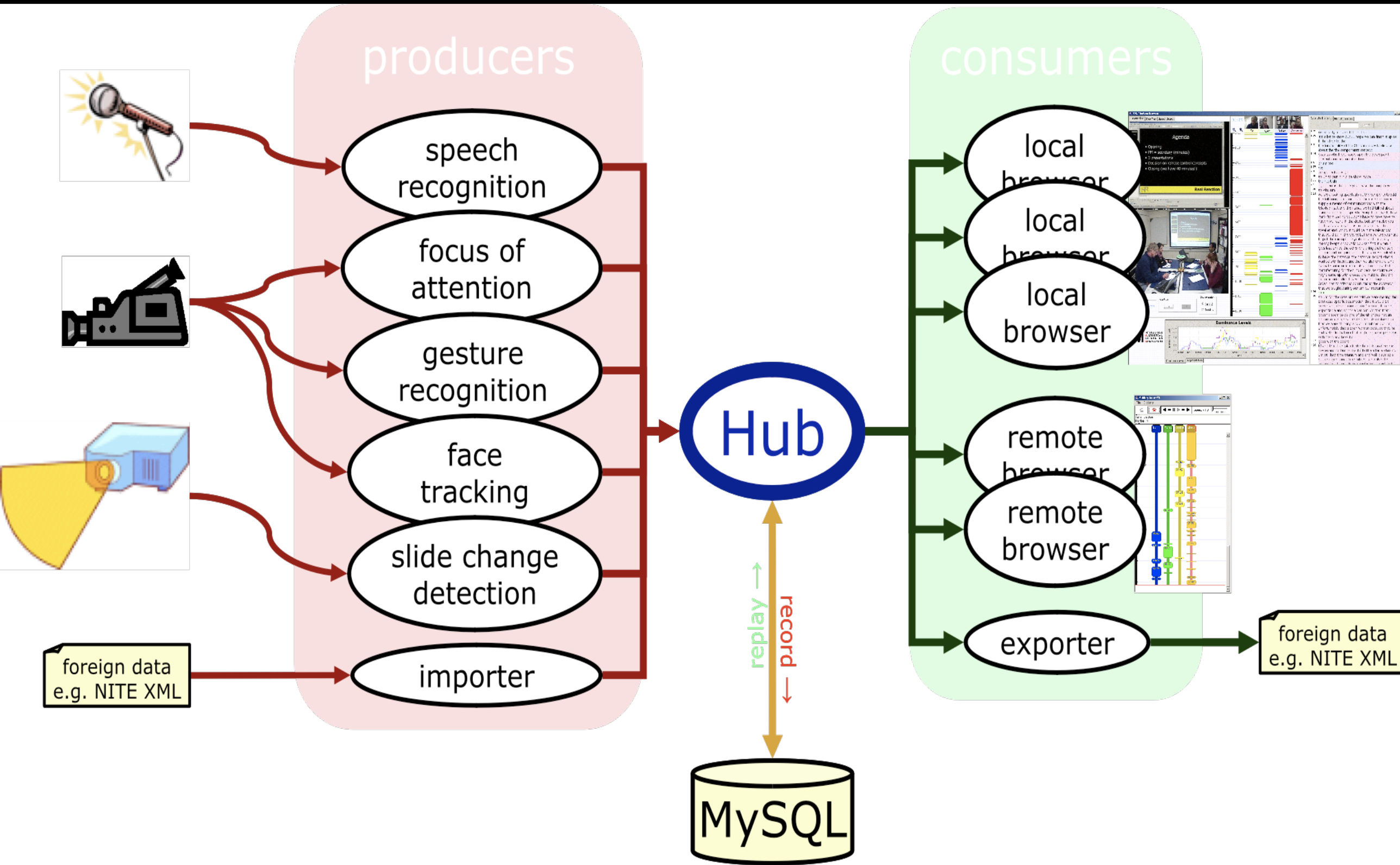
- Finding factors leading to a decision is a challenging task for users
- Automatic summaries outperform keyword spotting baselines
- Summaries of speech recognition transcripts
  - lower user satisfaction
  - perform the task almost as well as on human transcripts

Applications

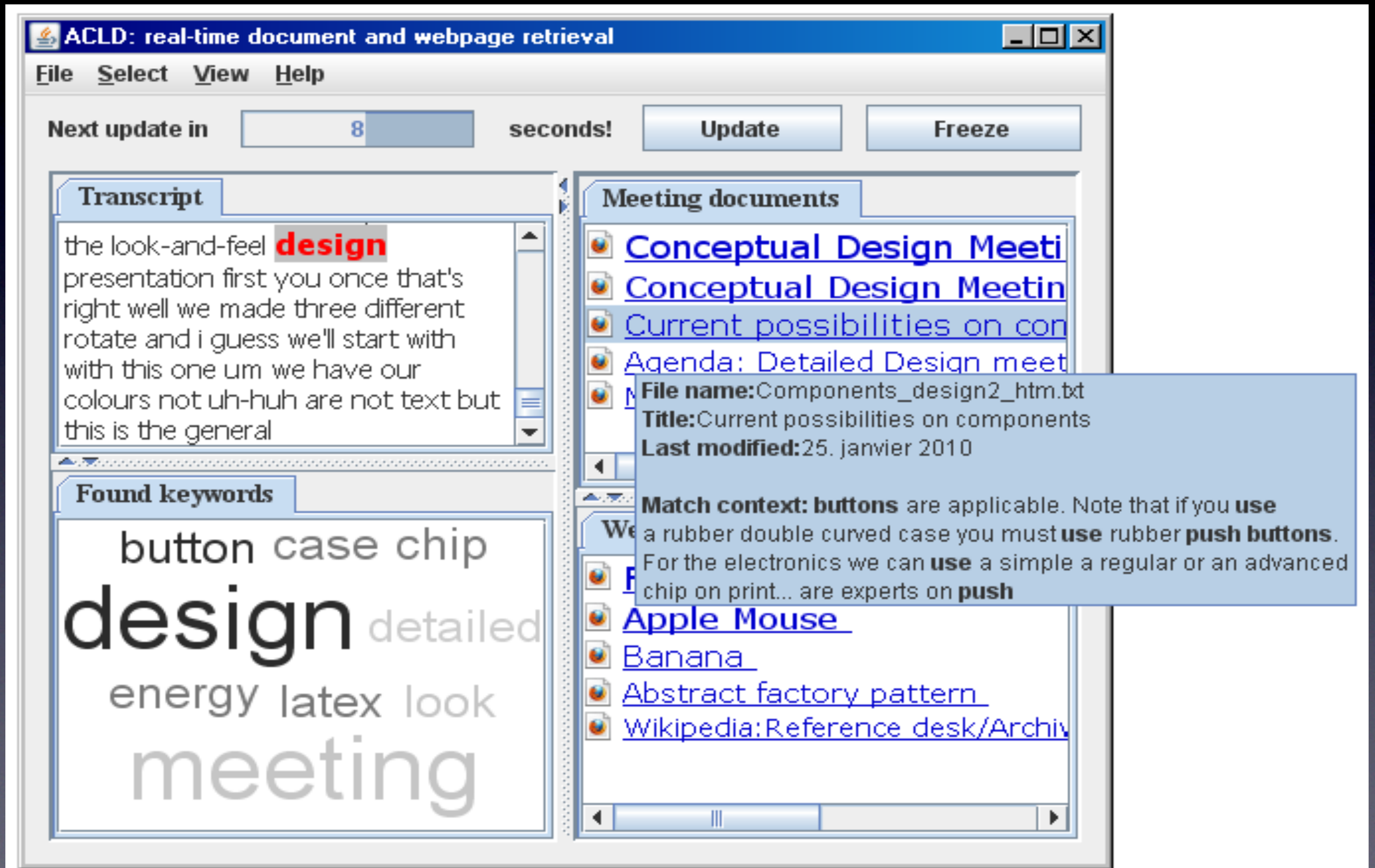
Browsing a recording



The AMI Hub


producers
speech recognition
focus of attention
gesture recognition
face tracking
slide change detection
importer
foreign data e.g. NITE XML
Hub
replay
record
MySQL
consumers
local browser
local browser
local browser
remote browser
remote browser
exporter
foreign data e.g. NITE XML

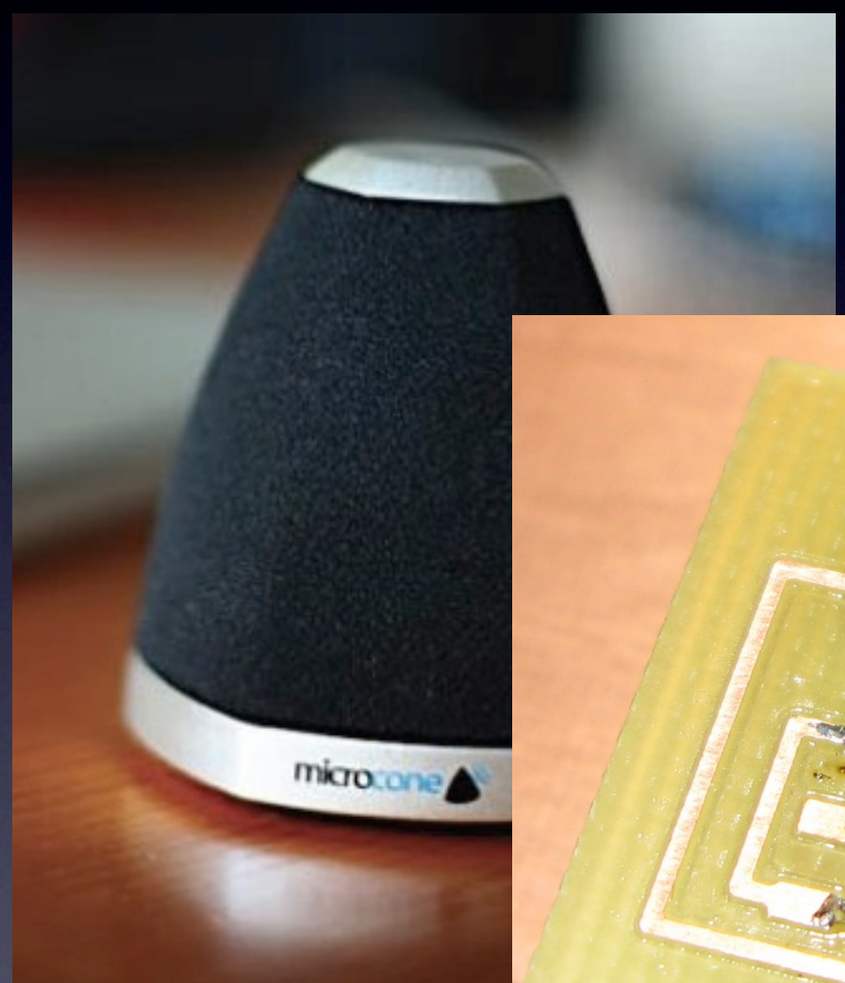
Content linking

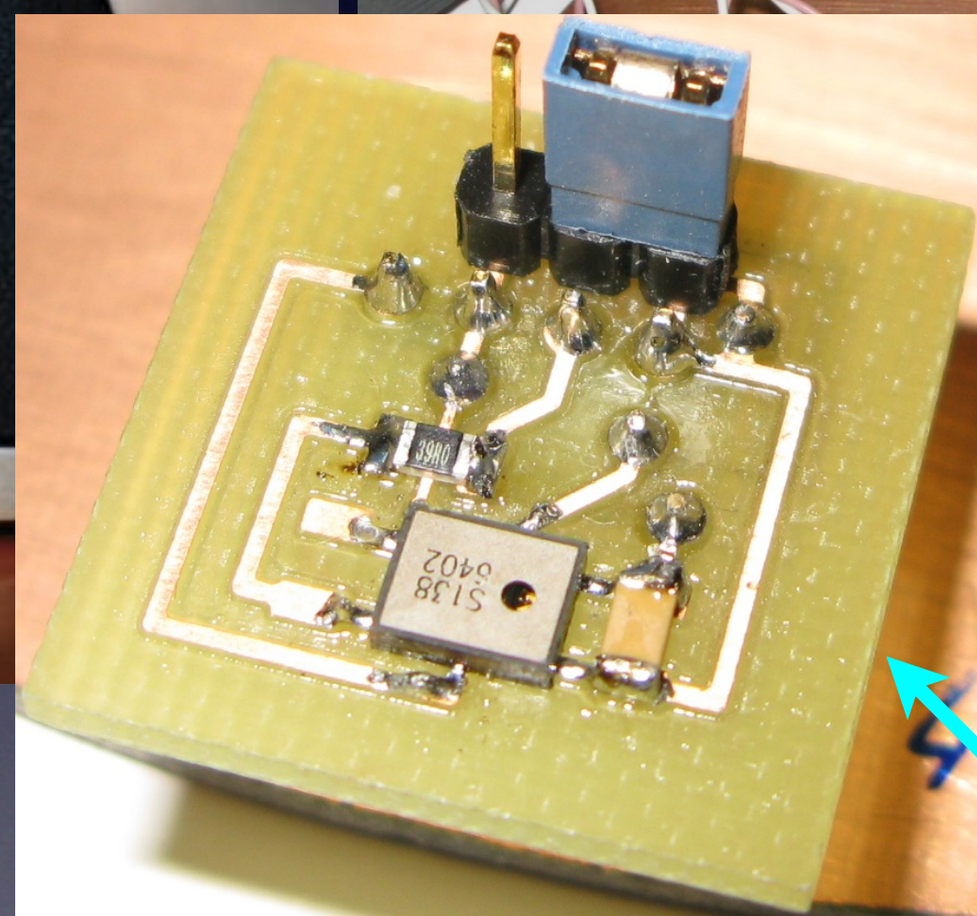


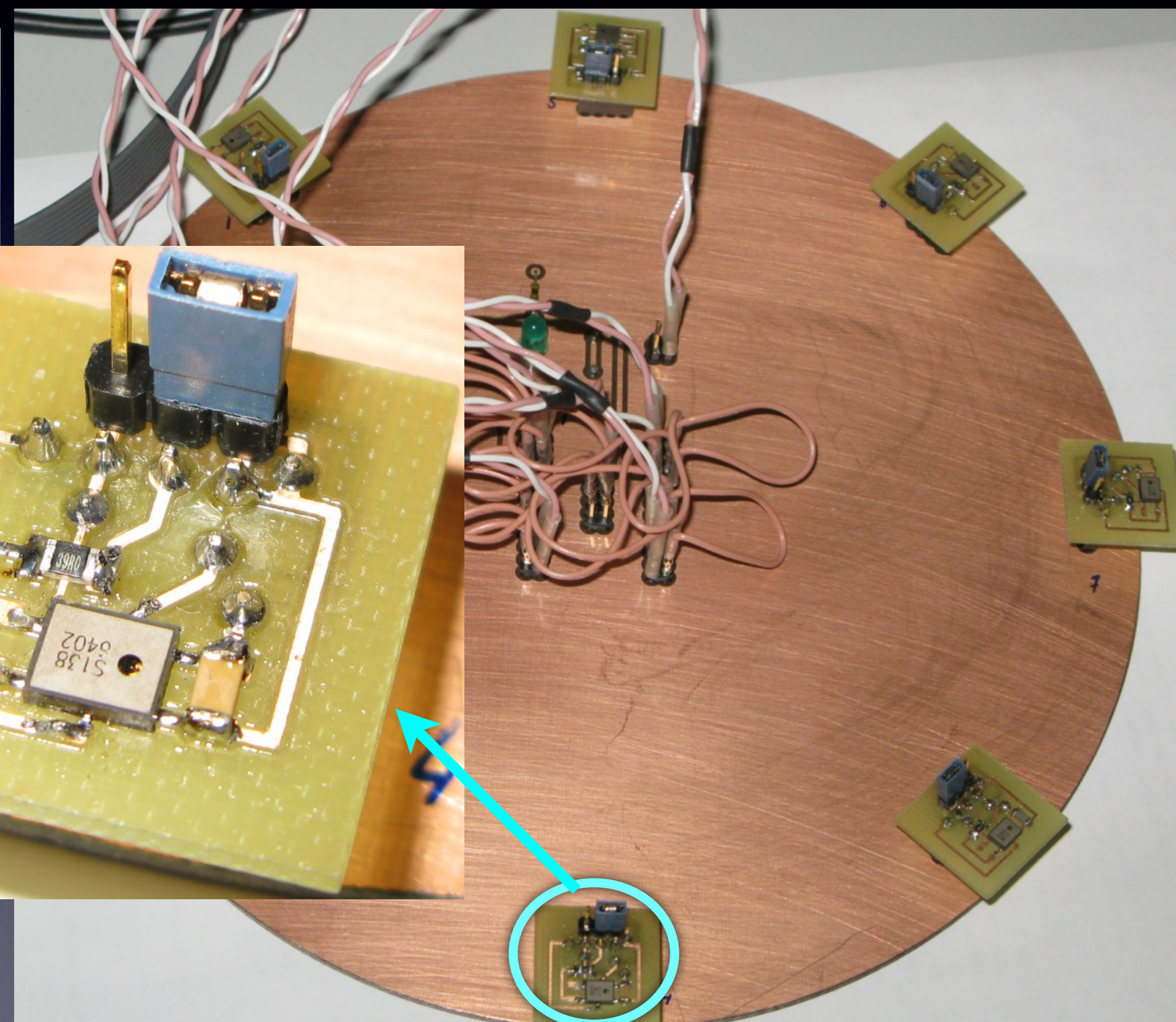
Meeting recording (c. 2005)

Meeting recording (2010)

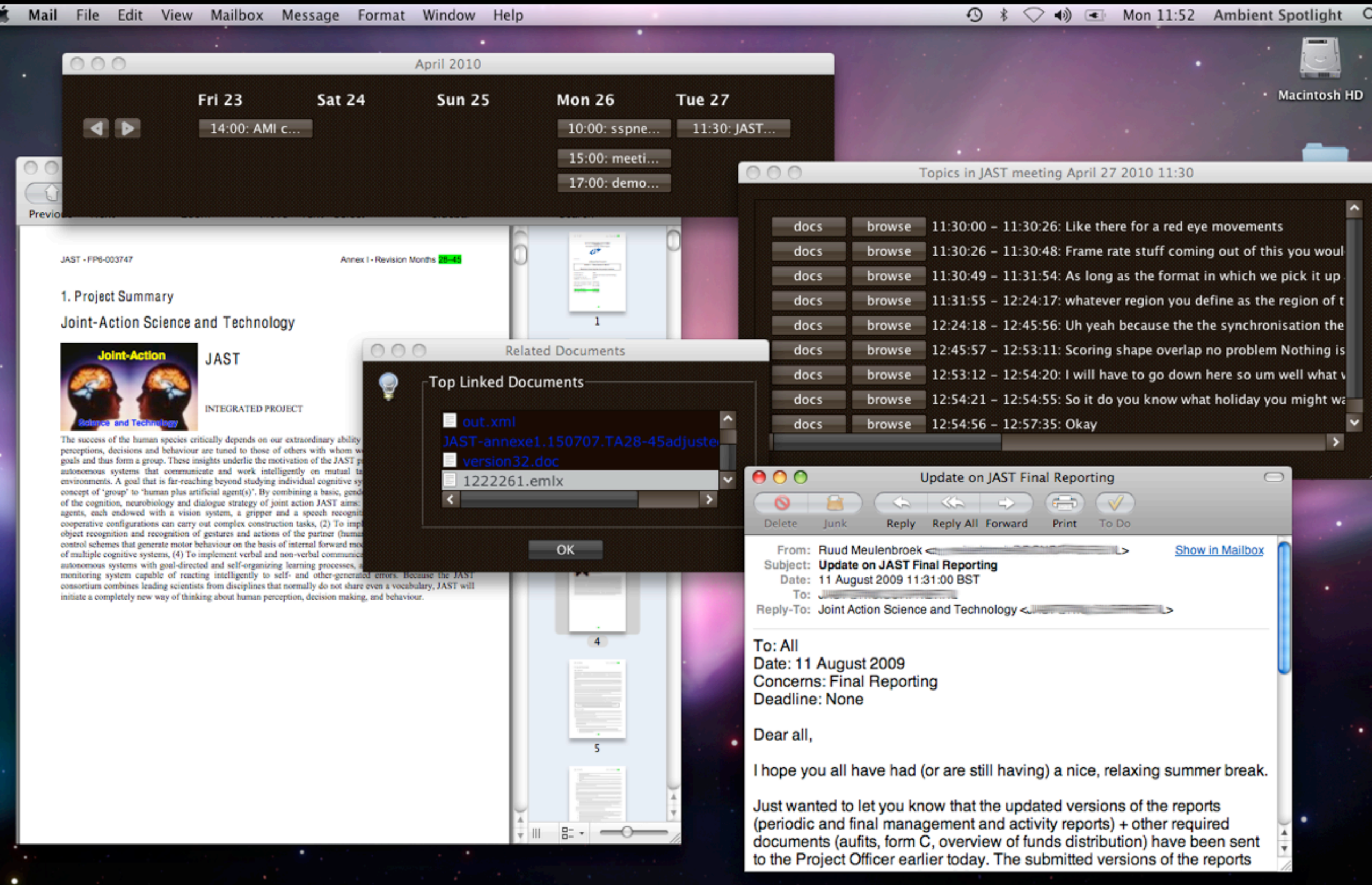
Commodity mic arrays







Ambient spotlight



Conclusions

- The AMI corpus is a great resource http://corpus.amiproject.org
- Combining multiple features / models is important
- Meeting speech recognition - high WERs, we need yet more advances in signal processing, acoustic modelling, language modelling
- Meeting interpretation - ASR transcripts, but also prosody, turn taking, focus of attention, ...
- Possible to build useful applications based on meeting analysis, recognition, and interpretation

Challenges

- Dealing with data from natural communication environments: multisource / multimodal / multiparty
- Adaptation, unsupervised learning
- Privacy and security
- Social aspects of communication
- Improve meetings, especially remote
- Lower error rates! (meaningful objective evaluations)

Thank you.

Acknowledgements to

- Long-time AMIs:

  Hervé Bourlard, Jean Carletta, Thomas Hain,
  Jonathan Kilgour, Mike Lincoln, Andrei Popescu-Belis

- Some current and former PhD students:

  Alfred Dielmann, Giulia Garau, Songfang Huang,
  Gabriel Murray, Le Zhang, Erich Zwyssig