

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



University of Twente
Enschede - The Netherlands



Noldus
Information Technology



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

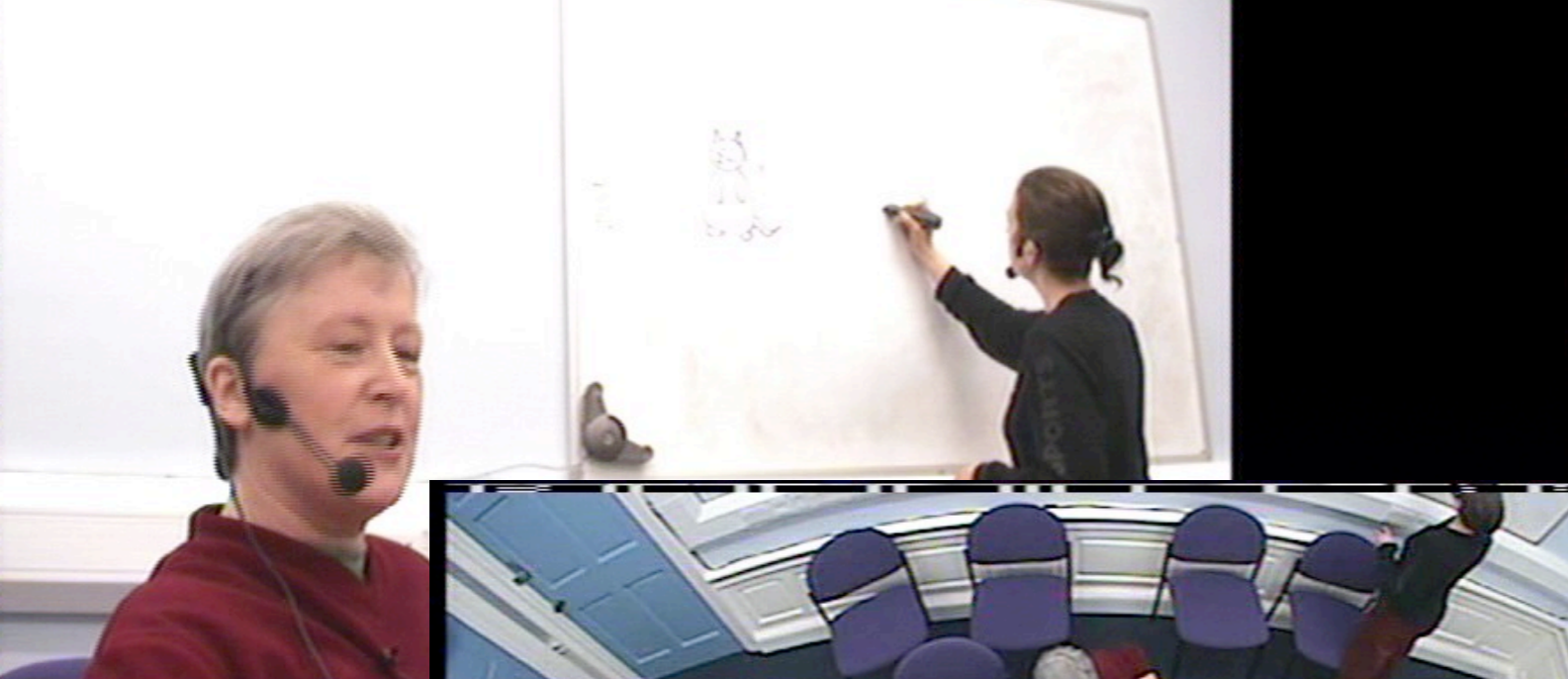
Recording multiparty interaction

- 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

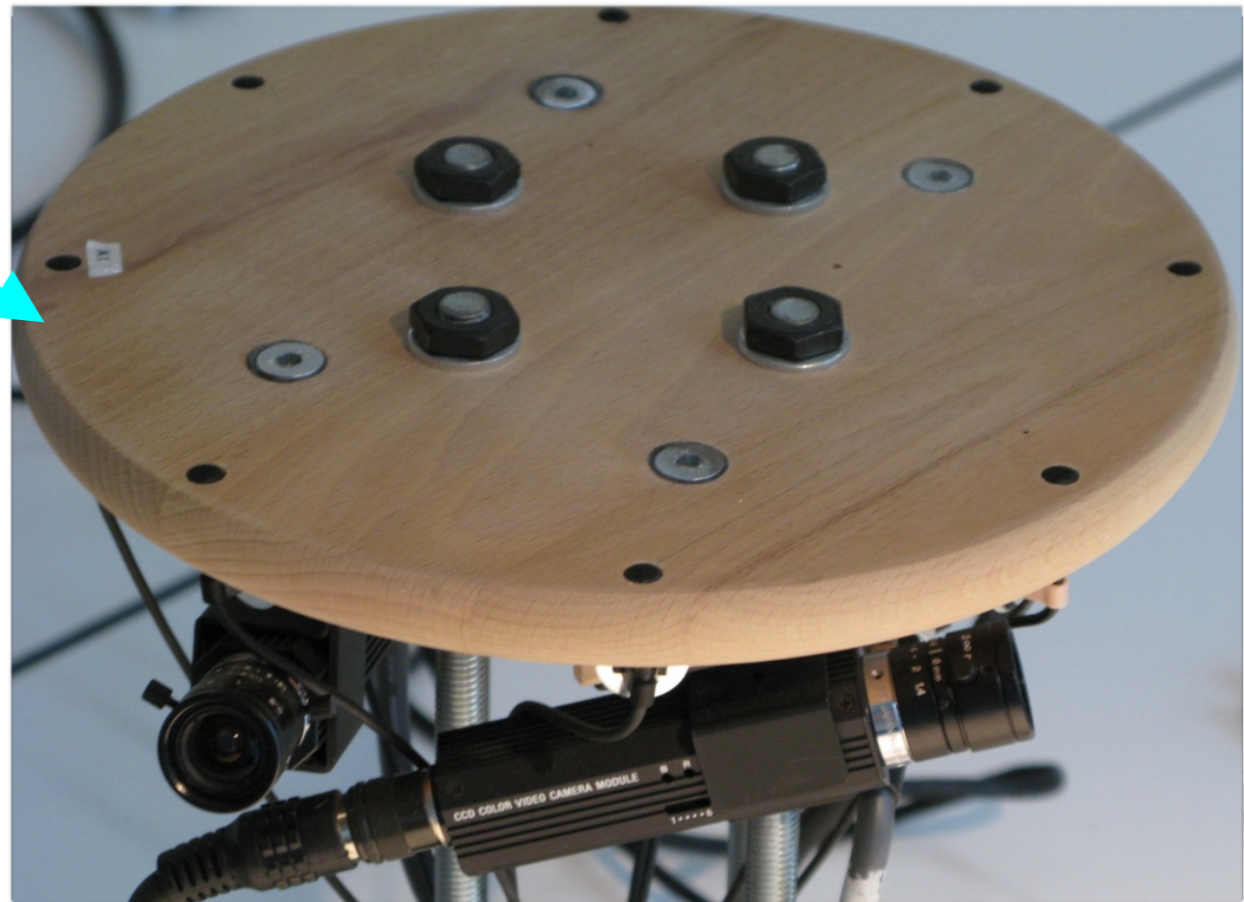


Mic Array

Cameras



Camera



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

The screenshot displays the 'Continuous Signal Labeler' application window. It features a menu bar with 'File', 'Annotate', and 'View'. The main interface is divided into several panels:

- NITE Video player:** Shows an overhead view of a meeting room with people around a table. A 'Mute' checkbox and a 'New' button are visible at the bottom.
- NITE Clock:** Includes playback controls (play, stop, next), a 'Sync Text Areas' checkbox, a time display of '0:07:55', a 'skip: 5' field, a 'Rate' slider with options from -4x to +4x, and a 'Reset' button.
- A - action-layer:** A text area for labeling video segments. It contains several lines of text with colored brackets indicating actions. The current segment is highlighted in blue and contains the text: `[moving location.....] [no action UNFINISHED]`.
- Legend:** A vertical stack of colored bars with corresponding labels: 'typed e - entering' (red), 'typed l - leaving' (green), 'typed u - standing up' (blue), 'typed d - sitting down' (yellow), 'typed m - moving location' (cyan), 'typed U - unconsidered action' (magenta), and 'typed n - no action' (grey).

Buttons for 'Delete', 'Edit Comment', and 'Finish' are located below the action-layer text area.

Dialogue act labelling

The screenshot displays the 'AMI Dialogue act coder' application. It features three main panels:

- Transcription:** Shows a dialogue transcript with speaker labels and timestamps. A context menu is open over the text '<my name is Francina>', listing dialogue act categories such as Acknowledgment, Informs, Requests, Suggests, Assessments, Social-Affective Acts, and Unclassifiable.
- Edit Adjacency Pairs:** A configuration window for an adjacency pair. It shows a source utterance 'A: Everybody ready' with the act 'Request Support' and a target utterance 'D: I think so' with the act 'Inform'. The type is set to 'POS'.
- Adjacency Pairs:** A list view showing two pairs. The second pair is highlighted, showing the source act 'Request Support' and target act 'Inform'.
- Edit Dialogue Acts:** A configuration window for a specific dialogue act. The agent is 'B', the DA type is '<none>', and the DA text is 'my name is Francina'. It includes options for addresssee, reflexivity, and buttons for 'Type...', 'Range...', 'All', 'Set Comment', 'New DA', and 'Delete!'.

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



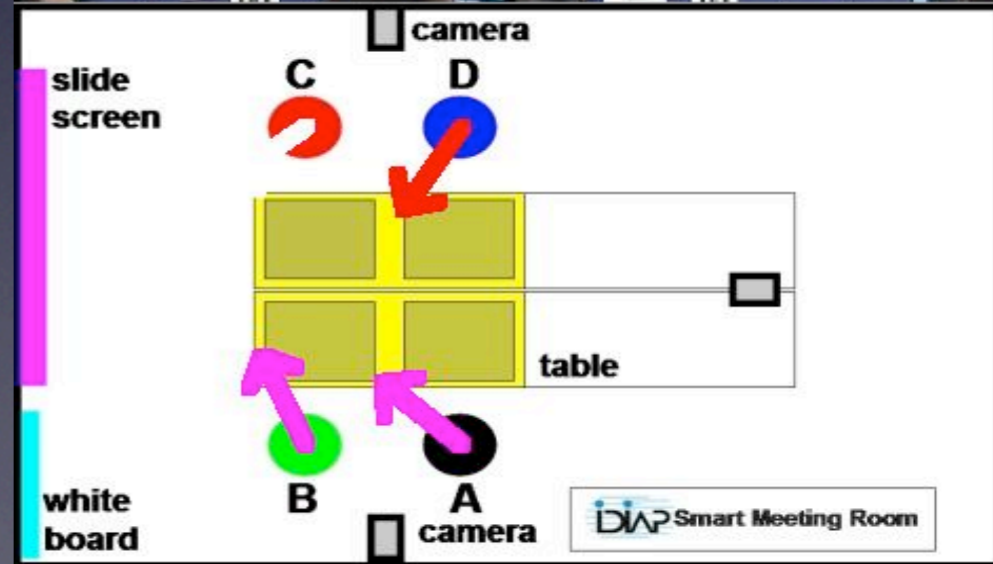
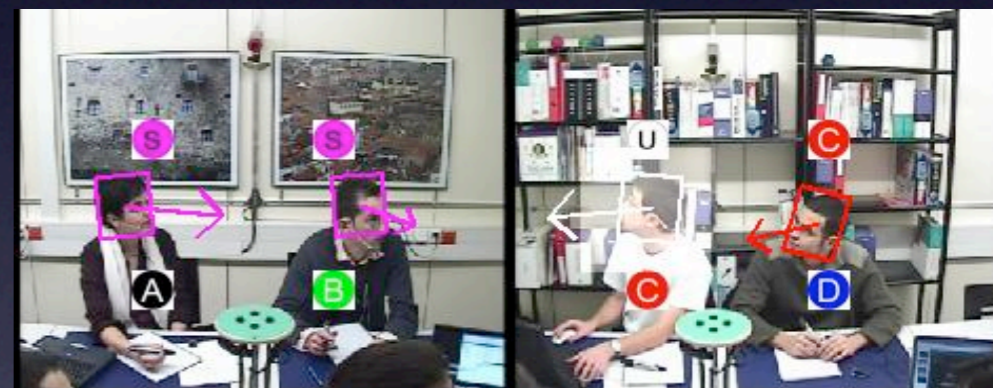
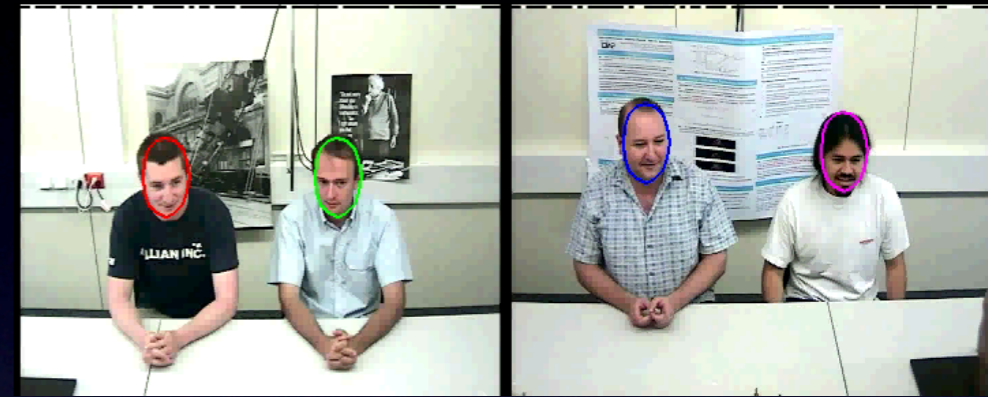
Gesturing while speaking



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

hmm

... so you have your energy source your user interface who's controlling the chip ...

click

rustle



“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

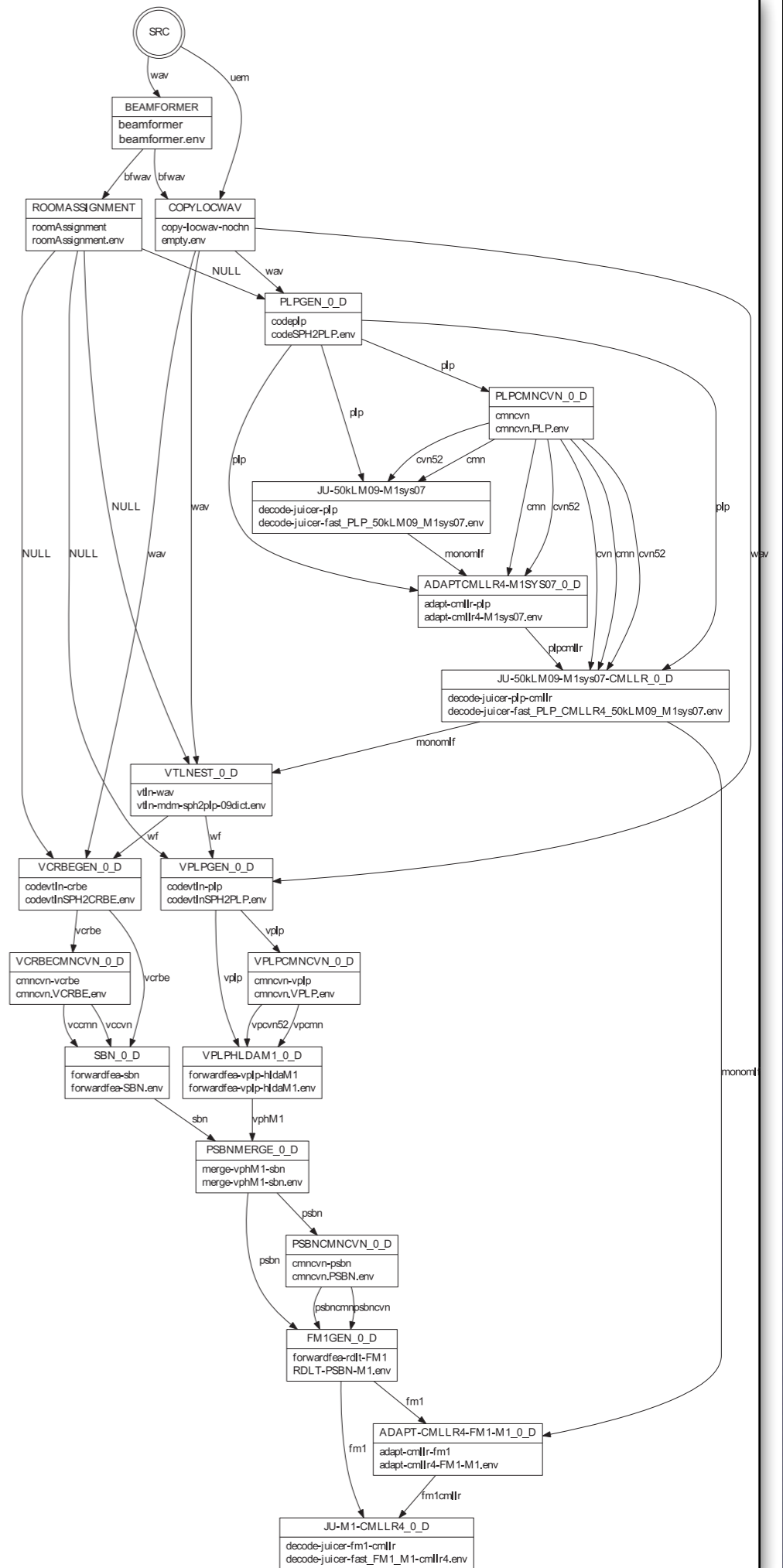
AMI-ASR



- 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

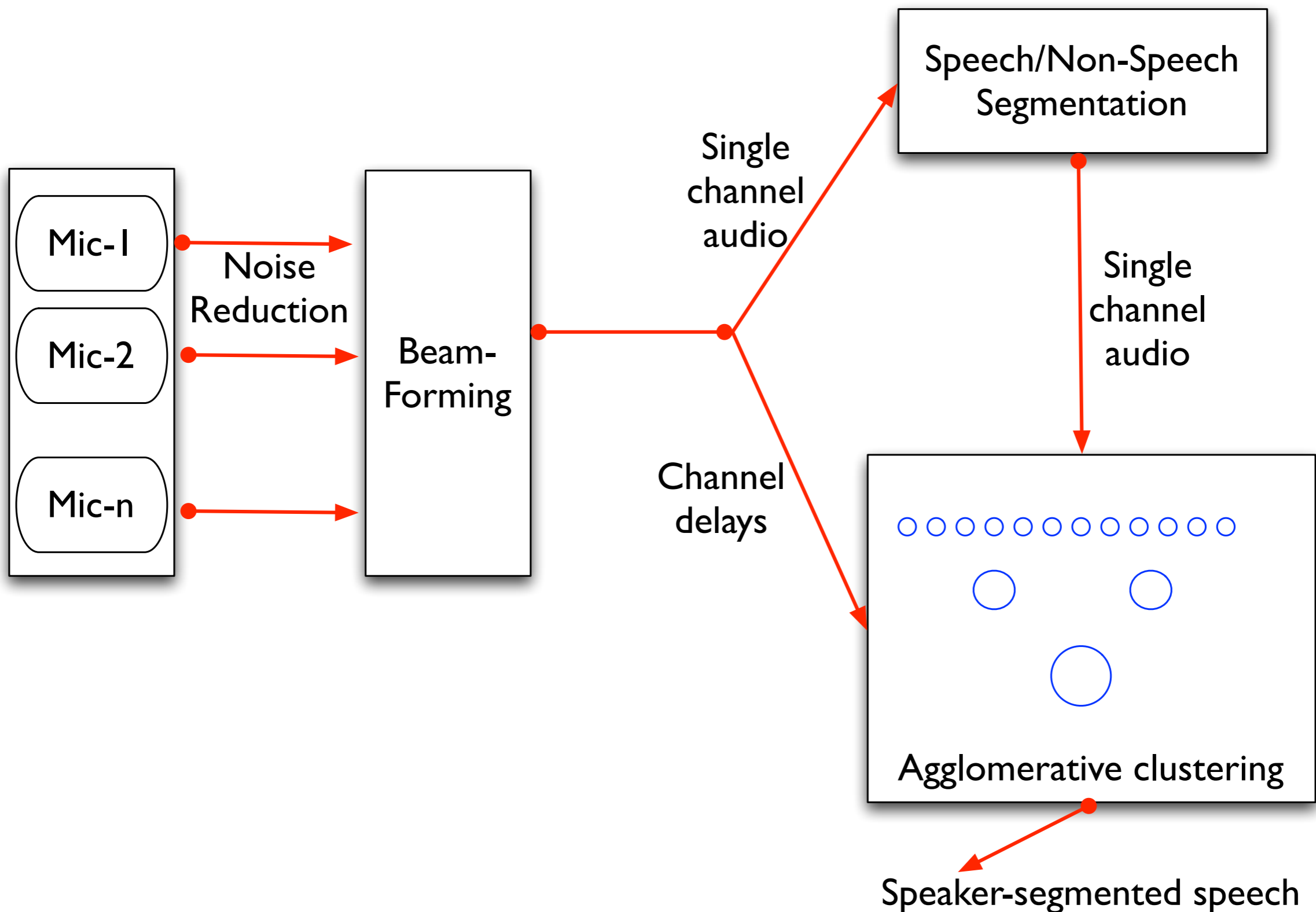
- Speech/non-speech segmentation
- PLP/MFCC features
- ML trained HMM/GMM system (122k 39D Gaussians)
- 50k vocabulary
- Trigram language model
- Weighted FST decoder



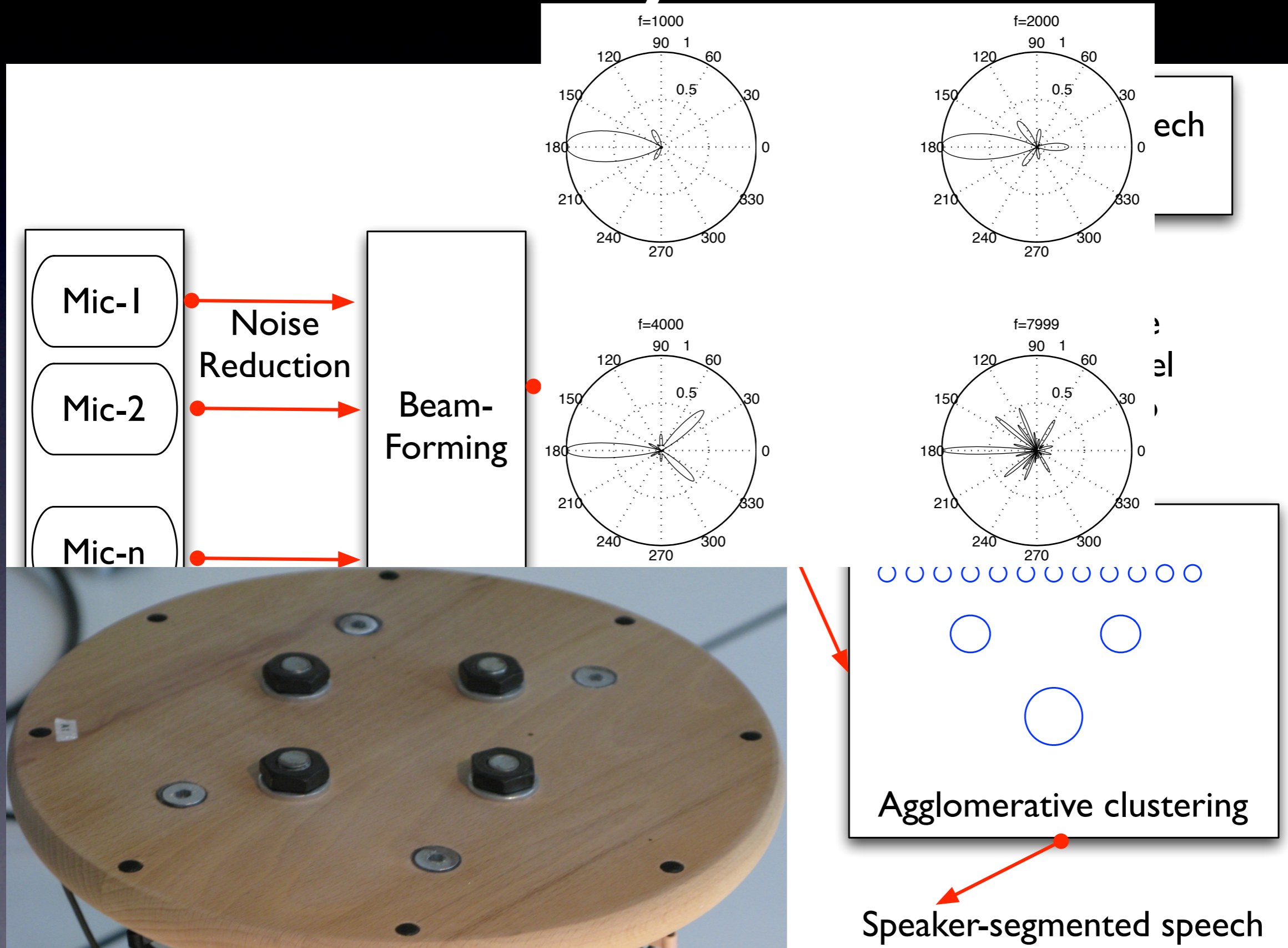
Additional components

- 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



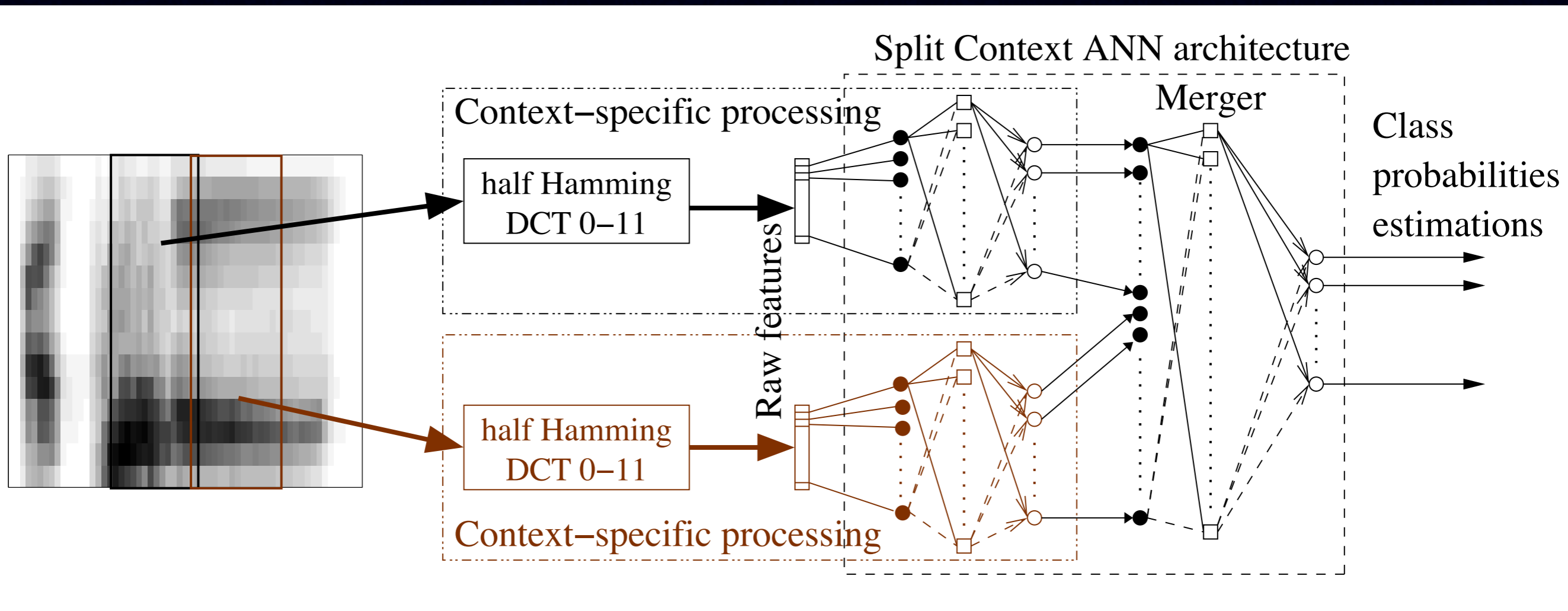
Mic array frontend



Discriminative features

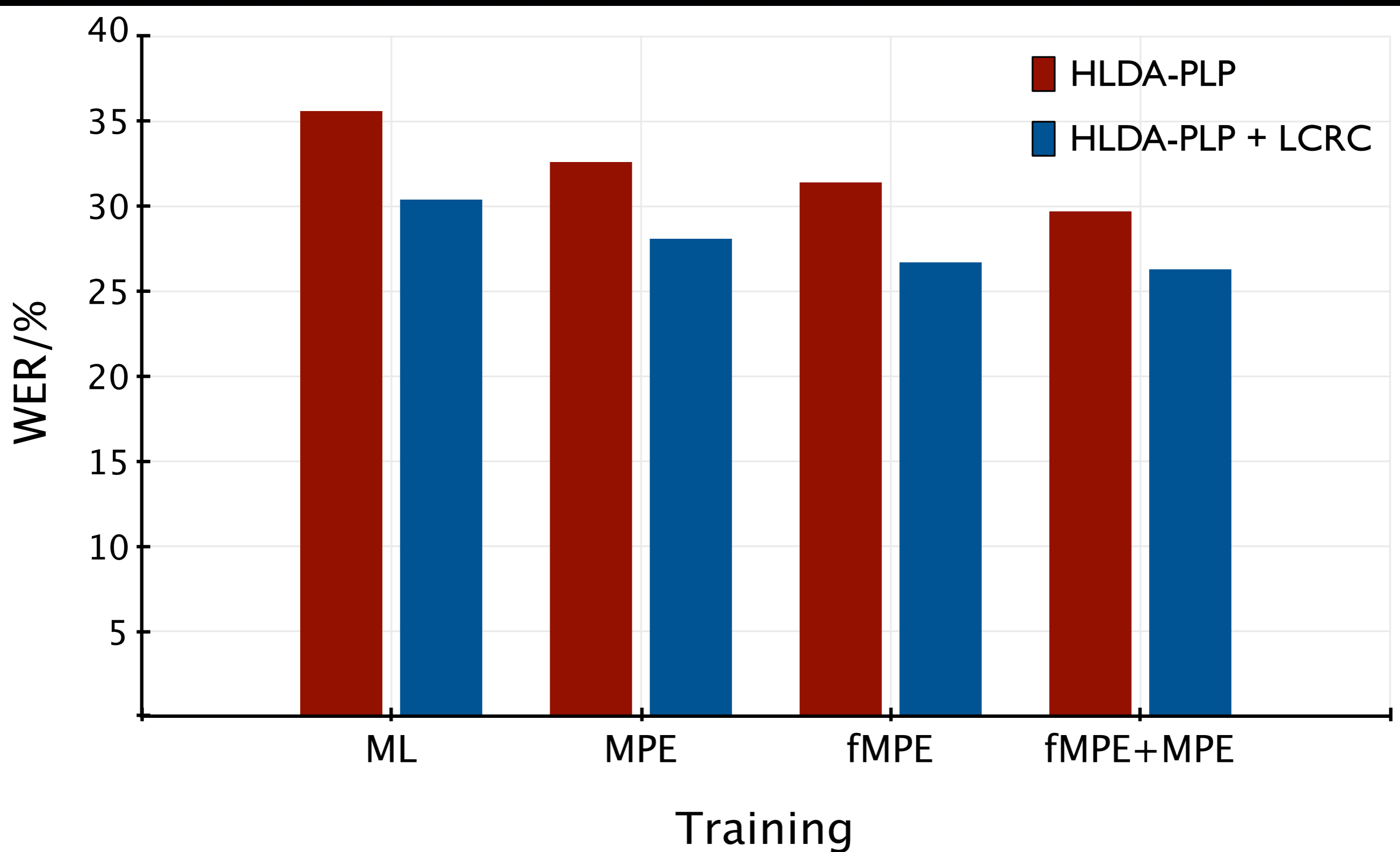
- Intensive amount of work on discriminative feature transforms (eg (H)LDA, fMPE)
- Posterior-based features from MLP phone classifiers
- Use as an additional feature stream
- Advantages
 - temporal context (± 25 frames)
 - encode phone discrimination information
 - weakly correlated with PLP/MFCC features

LCRC features

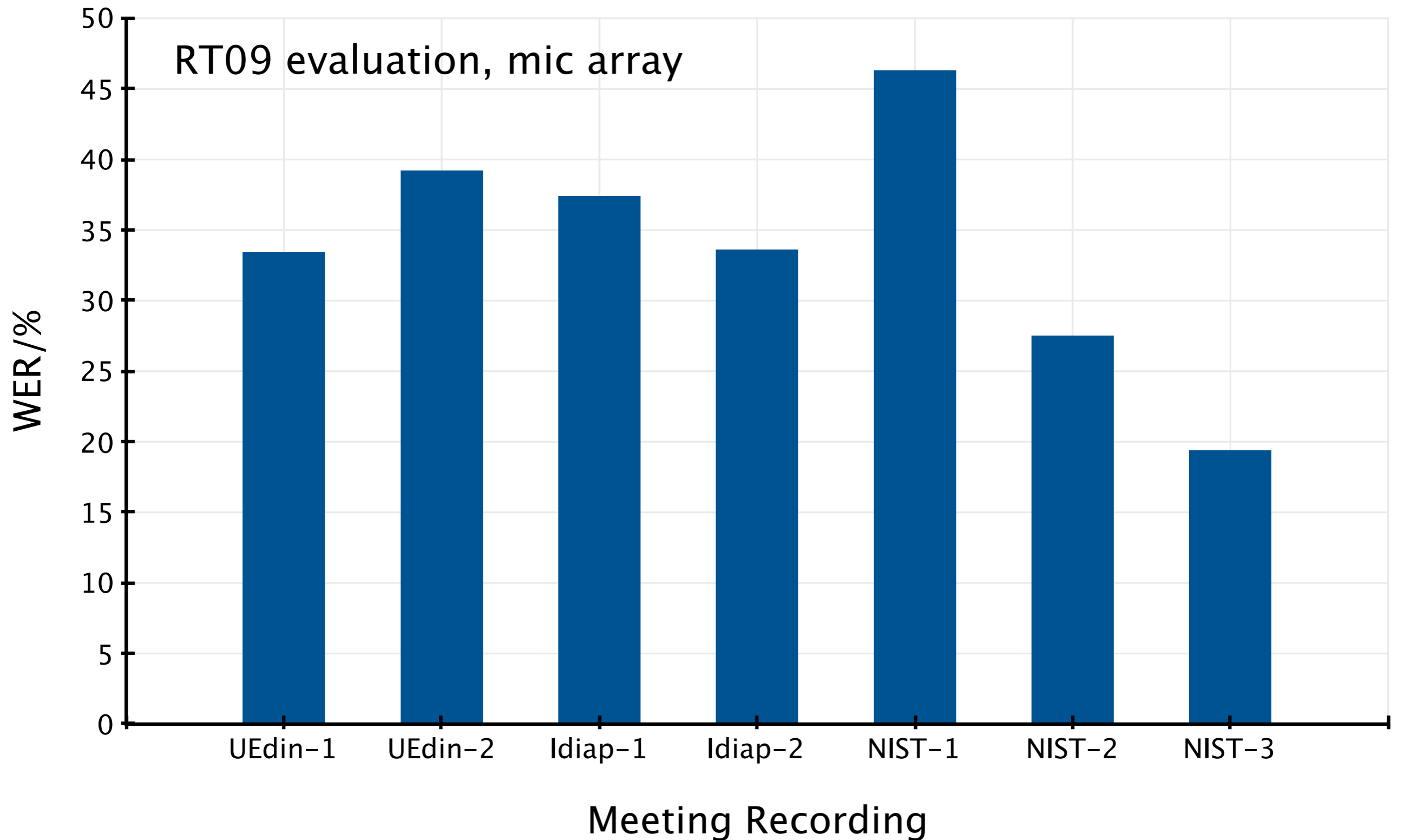


Karafiat, Grezl, Burget, Cernocky

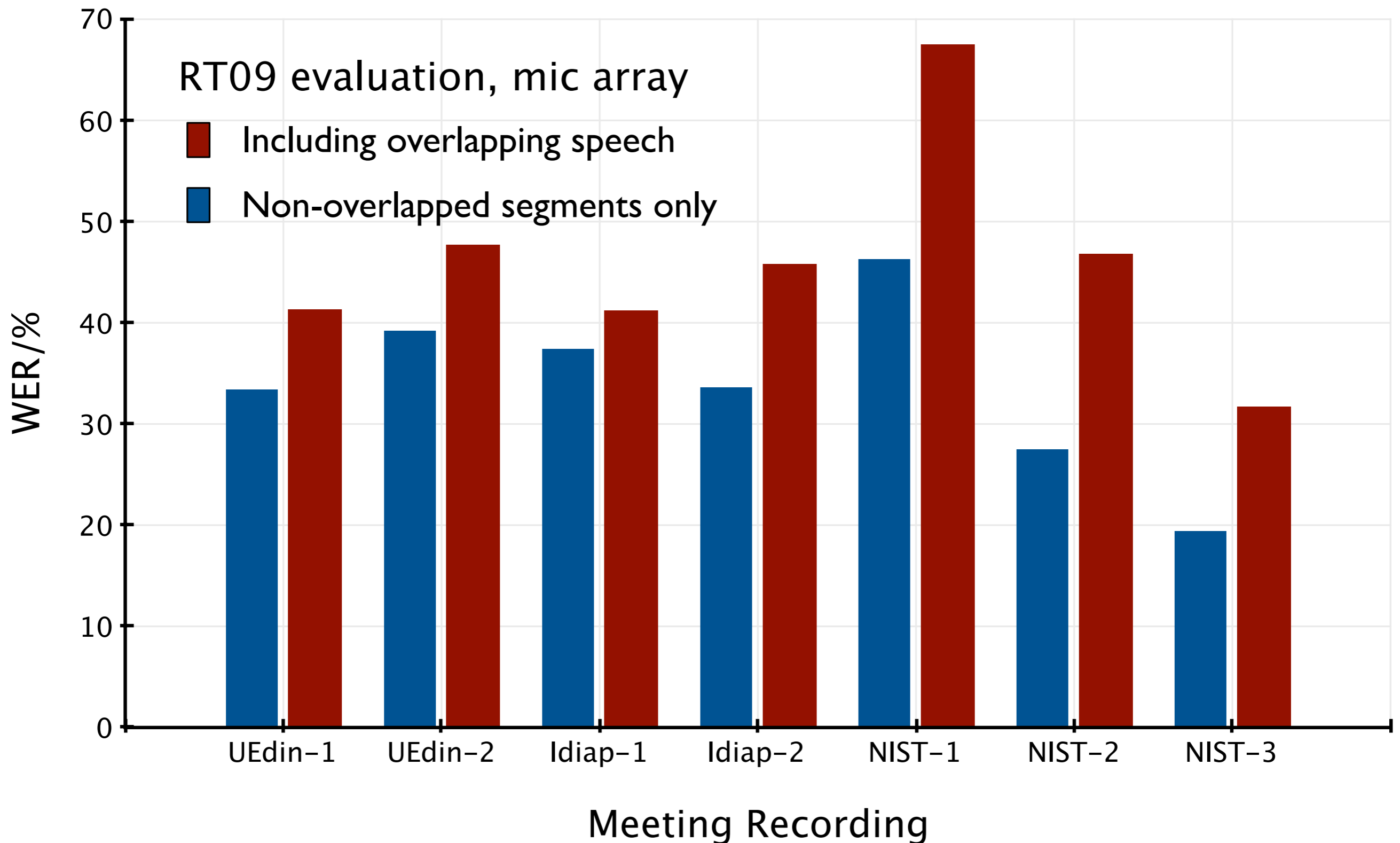
Results (RT07, IHM)



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

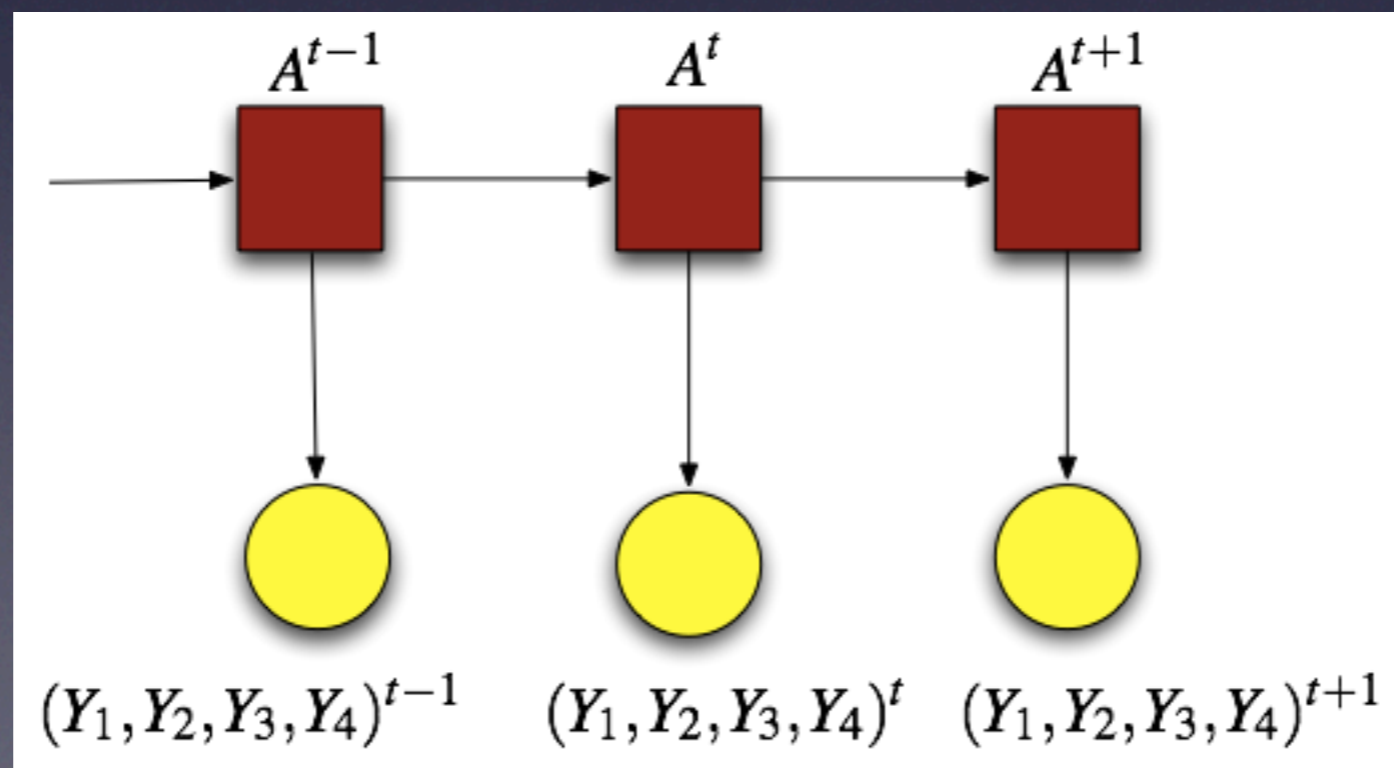
- Combine feature streams (speech, video, handwriting) to predict events in meetings
- Pilot study detection of meeting actions from a set of recorded meetings (M4 project)
 - Monologue
 - Discussion
 - Presentation
 - Speaking at whiteboard
 - Notetaking

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



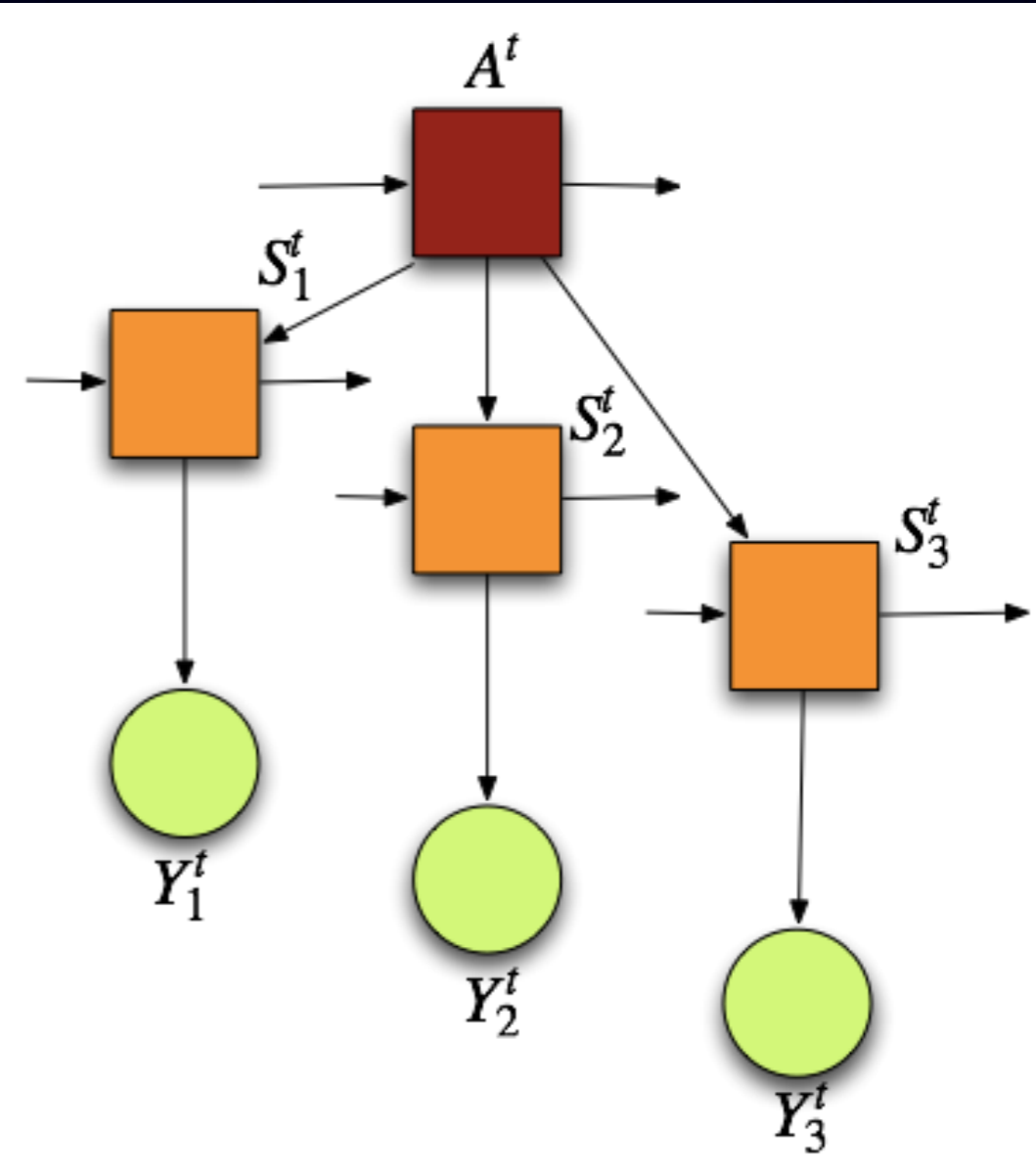
Baseline results

- 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

The screenshot displays the JFerret software interface, which is used for managing experiments and viewing transcriptions. The interface is divided into several sections:

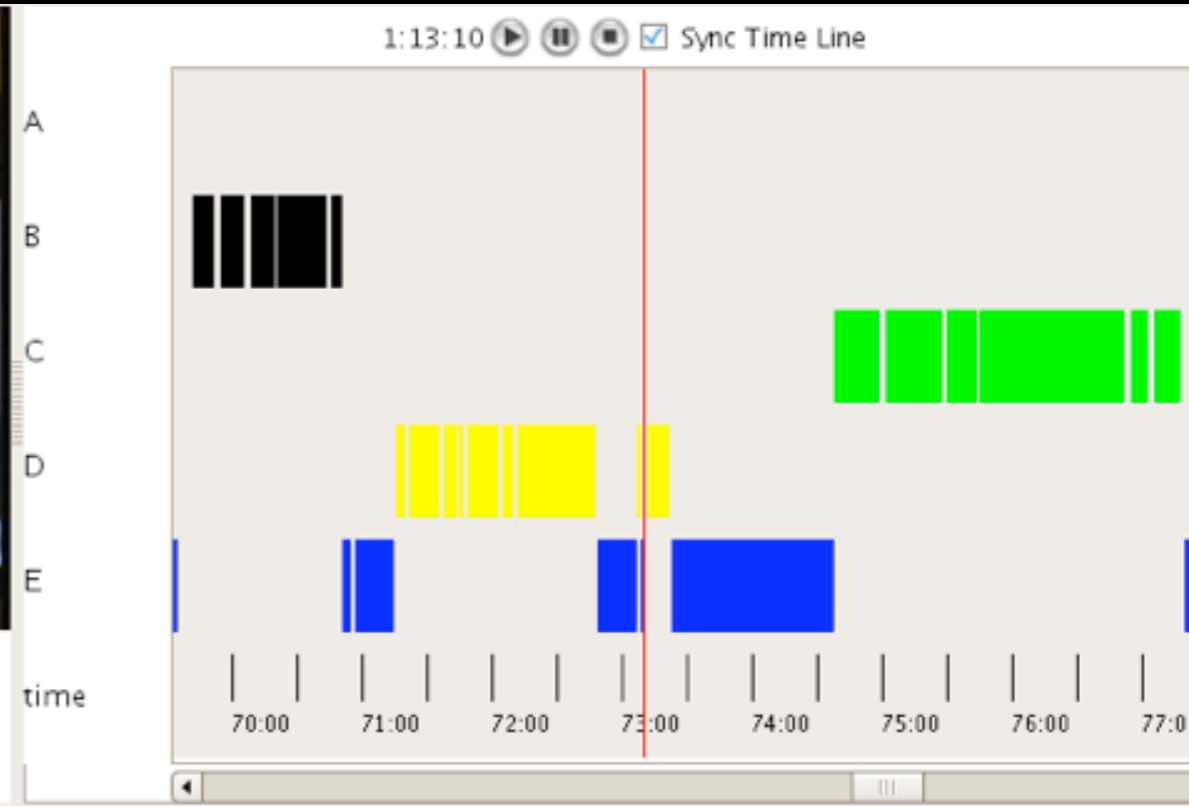
- Experiment control:** Located at the top, it contains two buttons: "Start Experiment" and "End Experiment".
- Tabbed interface:** Below the control section, there are several tabs labeled "AMI Meeting ES2008a", "AMI Meeting ES2008b", "AMI Meeting ES2008c", "AMI Meeting ES2008d", and "Typing tab".
- Video conference:** The central part of the interface shows a video conference with four participants in separate windows. The participants are wearing headsets and appear to be in a meeting room.
- media control:** Below the video windows, there is a media control bar with a timer showing "0:11:28" and icons for play, pause, and stop.
- ES2008a transcription:** On the bottom left, there is a transcription window for "ES2008a". It contains a list of dialogue turns with speaker labels (UI, ME, PM, ID) and their respective utterances. A yellow highlight is present on a specific line: "PM: Um . Okay , it seems we have a little bit of a conflict over um to uh combining all the remotes cont together versus having f five different remotes ; So um like you said you don't like having all the buttons on one on one remote , and yet you don't wanna have five remotes . So how do we work with that ?".
- extractive summary:** On the bottom right, there is a window titled "extractive summary" which displays a condensed version of the transcription. It lists key points from the dialogue, such as "UI - especially , like you know if I'm watching T_V_ I have have to have three separate" and "PM - it seems we have a little bit of a conflict over um to uh combining all the remote".

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



- Go touch okay done
- Go the reason its a dark on the other side
- Go a little test kind of the way first
- Go because they swallow but will be recorded on this side
- Go not just start explaining on this pervasive computing
- Go that you cannot separate the the context in contents
- Go just before to meetings so its just
- Go itll virtual okay one of the powerful
- Go we started to work with them and we

Sync Transcript

Segment Excision

E: it is an a device for those and users but somebody has to create the content that has the one for the device and then find out the possibilities and um so here and probably making this device for both a bit the same time the so that's uh that's one

D: situations will be different rerun the one yeah..... is a student of your role..... so what advertisement mediocre big might be great solution that's.. so commented beyond that are you have an issue issue which is say you want to create a device that and rebels real world interactions with digital content and that suggests that the content that people interact with it'll virtual okay one of the powerful things that we just case is that the content also is in the real world not just oracle going to doesn't just use crystal as a way to navigate the database of historical curiosities but actually the places you are the town and the real things you see around you are connected to the things you get from the digital sources yes but i think the role of the real world is is larger than the issue because

E: it's not written and then an interface it is also point of the content but yes very but uh yes so i have the first interaction vision we discourage but i think we'll skip that because it's it is just what we we we've talked about um

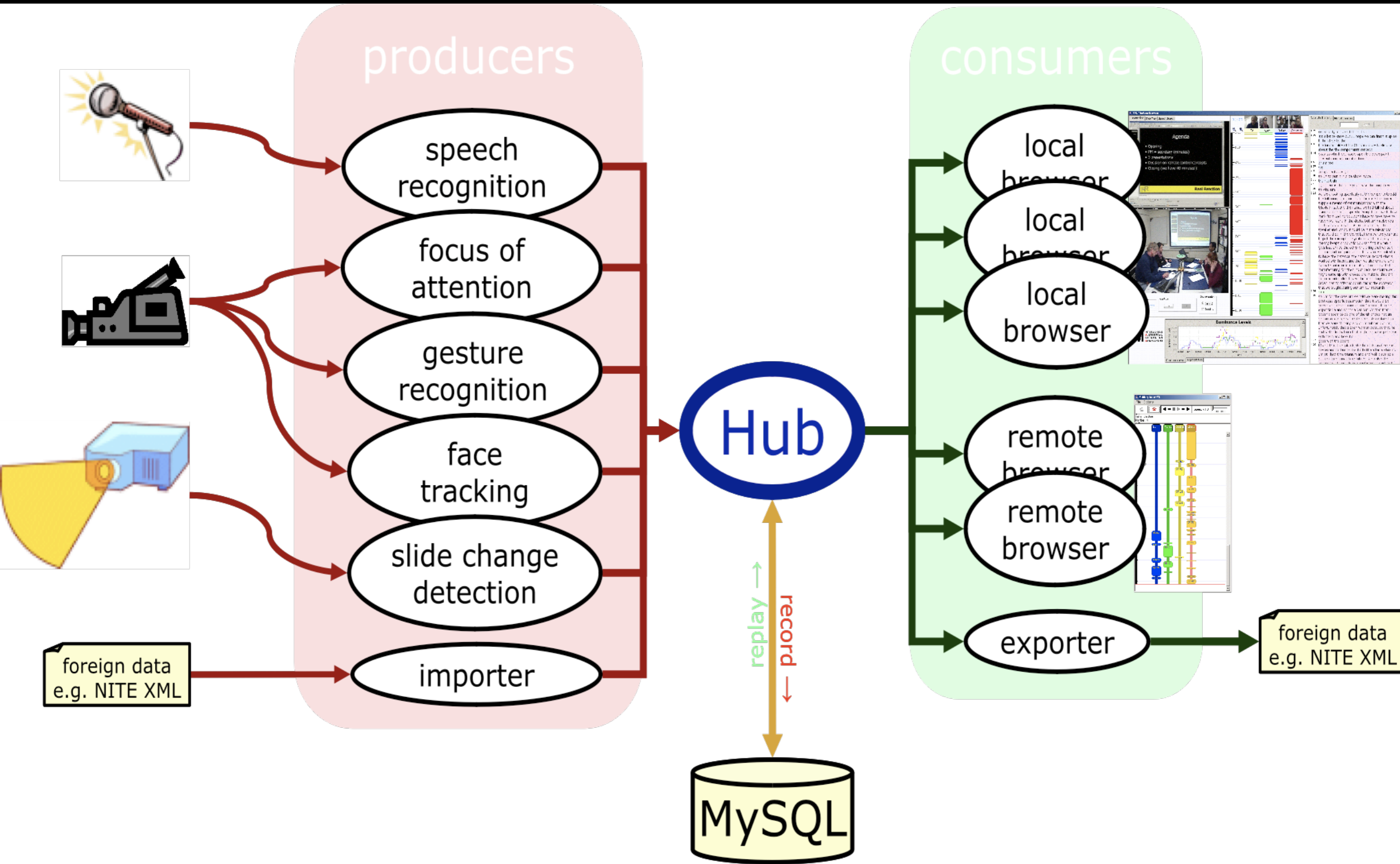
D:

E: !....

D: is maybe a context vision..

E: with the need any of opinion or feedback going to rewrite rewrite a prototype more examples in it was to uh already set the better structure nico incorporate feedback kind of two presented with a nice uh document i will clear a lot of the publication proposal we talked about um so refined is interaction vision and then by the hope to goal finally to the

The AMI Hub



Content linking

The screenshot displays the ACLD (real-time document and webpage retrieval) application window. The interface includes a menu bar (File, Select, View, Help), a status bar (Next update in 8 seconds!), and two main panes: a Transcript pane and a Meeting documents pane.

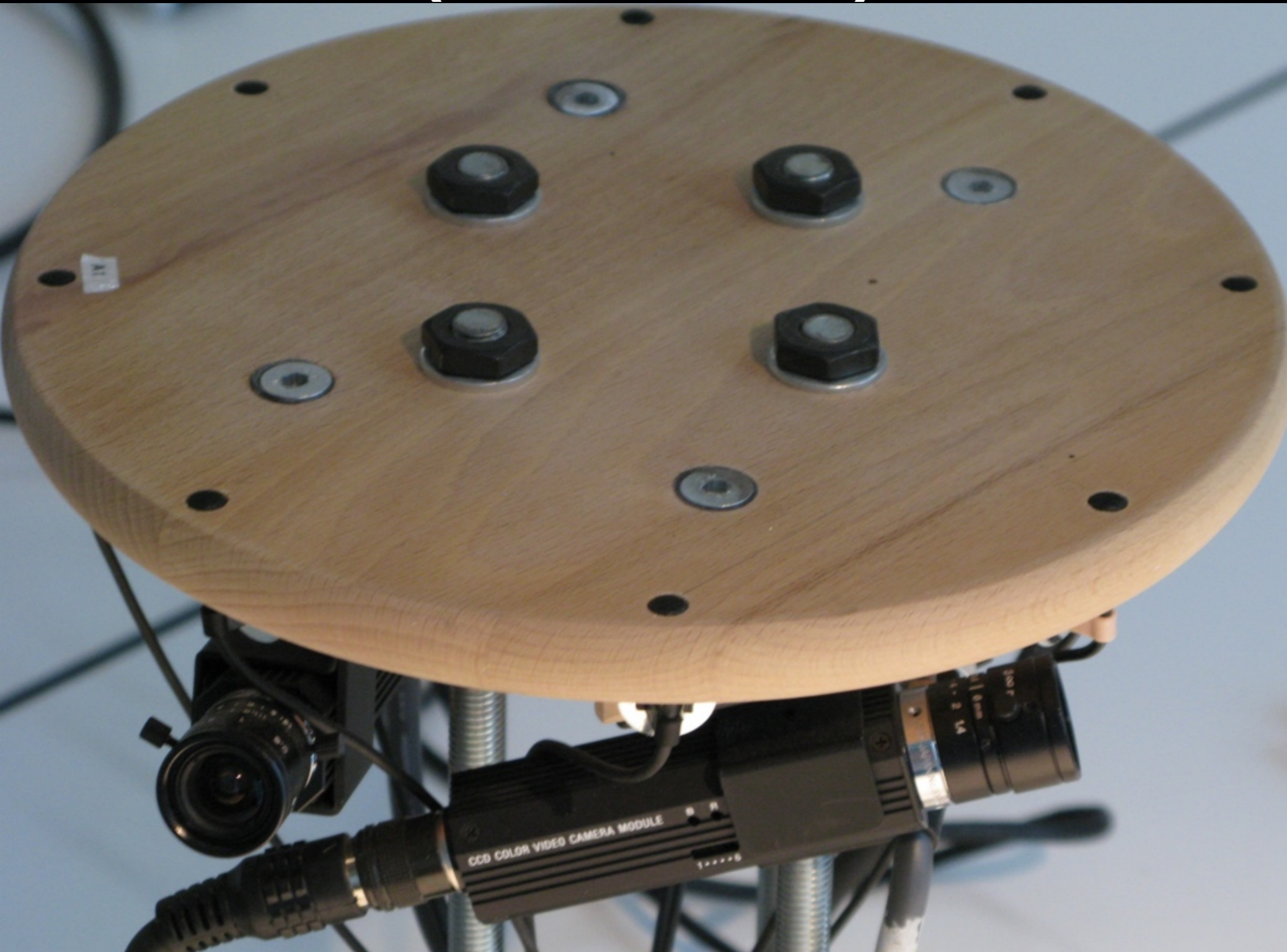
Transcript: The transcript shows a text snippet: "the look-and-feel **design** presentation first you once that's right well we made three different rotate and i guess we'll start with with this one um we have our colours not uh-huh are not text but this is the general". The word "design" is highlighted in red.

Found keywords: A list of keywords is displayed, including "button case chip", "design", "energy latex look", and "meeting". The word "design" is the largest and most prominent.

Meeting documents: A list of documents is shown, including "Conceptual Design Meeti", "Conceptual Design Meetin", "Current possibilities on con", "Agenda: Detailed Design meet", and "Apple Mouse". The document "Current possibilities on con" is selected, and a tooltip is displayed over it.

Tooltip: The tooltip provides metadata for the selected document: "File name: Components_design2_hm.txt", "Title: Current possibilities on components", and "Last modified: 25. janvier 2010". It also shows a snippet of the document's content: "Match context: buttons are applicable. Note that if you use a rubber double curved case you must use rubber push buttons. For the electronics we can use a simple a regular or an advanced chip on print... are experts on push".

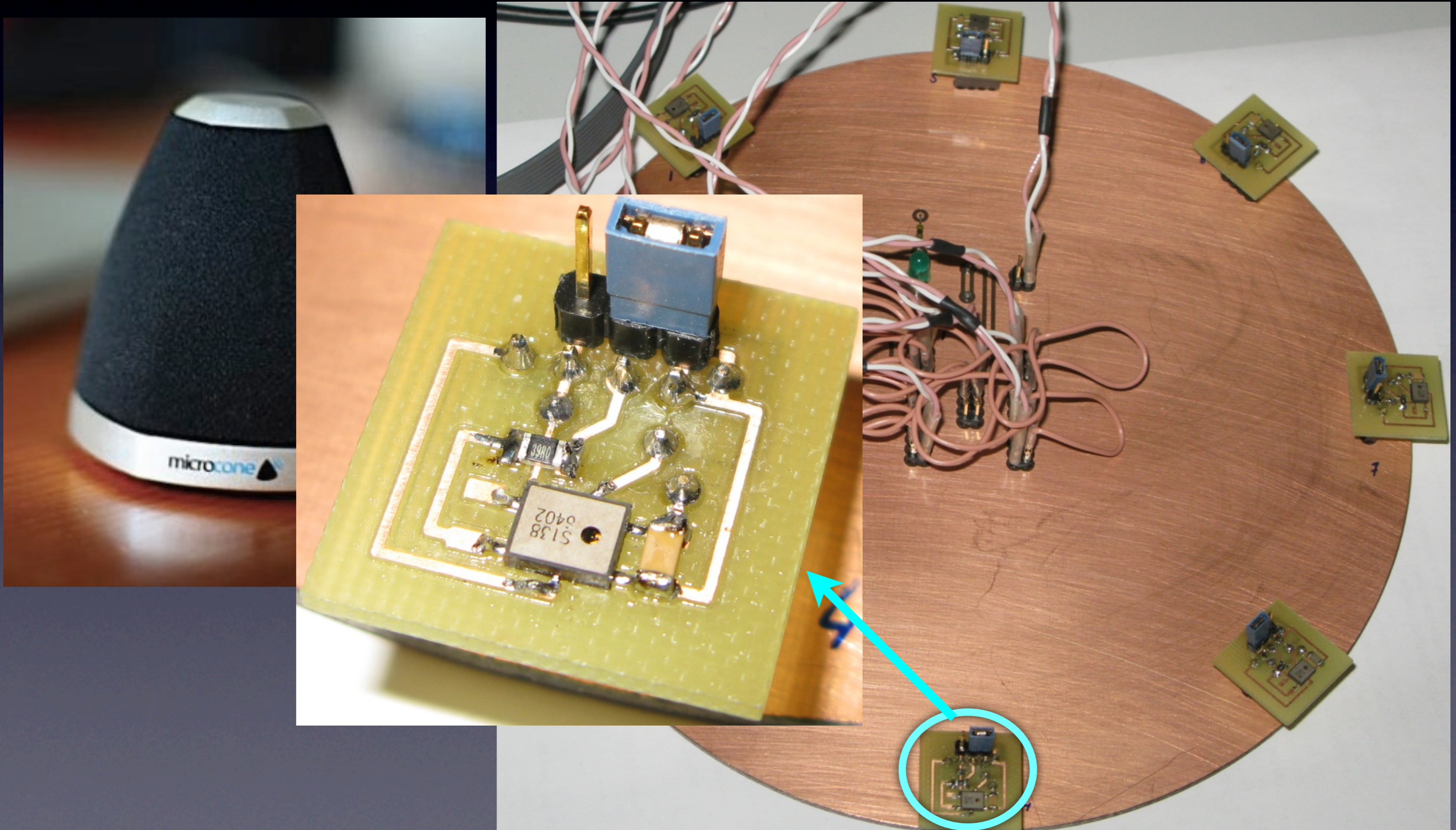
Meeting recording (c. 2005)



Meeting recording (2010)



Commodity mic arrays



Ambient spotlight

Mail File Edit View Mailbox Message Format Window Help Mon 11:52 Ambient Spotlight

April 2010

Fri 23 Sat 24 Sun 25 Mon 26 Tue 27

14:00: AMI c... 10:00: sspne... 11:30: JAST...
15:00: meeti...
17:00: demo...


Topics in JAST meeting April 27 2010 11:30

docs browse 11:30:00 - 11:30:26: Like there for a red eye movements
docs browse 11:30:26 - 11:30:48: Frame rate stuff coming out of this you woul
docs browse 11:30:49 - 11:31:54: As long as the format in which we pick it up
docs browse 11:31:55 - 12:24:17: whatever region you define as the region of t
docs browse 12:24:18 - 12:45:56: Uh yeah because the the synchronisation the
docs browse 12:45:57 - 12:53:11: Scoring shape overlap no problem Nothing is
docs browse 12:53:12 - 12:54:20: I will have to go down here so um well what v
docs browse 12:54:21 - 12:54:55: So it do you know what holiday you might wa
docs browse 12:54:56 - 12:57:35: Okay

JAST - FP6-003747 Annex I - Revision Months

1. Project Summary

Joint-Action Science and Technology



JAST
INTEGRATED PROJECT

The success of the human species critically depends on our extraordinary ability perceptions, decisions and behaviour are tuned to those of others with whom we share goals and thus form a group. These insights underlie the motivation of the JAST project to develop autonomous systems that communicate and work intelligently on mutual tasks in natural environments. A goal that is far-reaching beyond studying individual cognitive systems is the concept of "group" to "human plus artificial agent(s)". By combining a basic, general concept of the cognition, neurobiology and dialogue strategy of joint action JAST aims to: (1) To implement cooperative configurations can carry out complex construction tasks, (2) To implement object recognition and recognition of gestures and actions of the partner (human or robot), (3) To implement control schemes that generate motor behaviour on the basis of internal forward models of multiple cognitive systems, (4) To implement verbal and non-verbal communication in autonomous systems with goal-directed and self-organizing learning processes, and (5) To implement a monitoring system capable of reacting intelligently to self- and other-generated errors. Because the JAST consortium combines leading scientists from disciplines that normally do not share even a vocabulary, JAST will initiate a completely new way of thinking about human perception, decision making, and behaviour.

Related Documents

Top Linked Documents

- out.xml
- JAST-annexe1.150707.TA28-45adjuste
- version32.doc
- 1222261.emlx

Update on JAST Final Reporting

Delete Junk Reply Reply All Forward Print To Do

From: Ruud Meulenbroek
Subject: Update on JAST Final Reporting
Date: 11 August 2009 11:31:00 BST
To: Joint Action Science and Technology
Reply-To: Joint Action Science and Technology

To: All
Date: 11 August 2009
Concerns: Final Reporting
Deadline: None

Dear all,

I hope you all have had (or are still having) a nice, relaxing summer break.

Just wanted to let you know that the updated versions of the reports (periodic and final management and activity reports) + other required documents (aufits, form C, overview of funds distribution) have been sent to the Project Officer earlier today. The submitted versions of the reports

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 Boutilard, 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 Zwysig