Meeting Recognition

Thomas Hain
University of Sheffield
t.hain@dcs.shef.ac.uk

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk
Meeting Recognition

Thomas Hain
University of Sheffield
t.hain@dcs.shef.ac.uk

Steve Renals
University of Edinburgh
s.renals@ed.ac.uk

Centre for Speech Technology Research
We spend a lot of time in meetings
Hand-annotated meeting records are not as rich as we might like.

Some text remains on a held line:

Last output media
percent of on paper-based
character-based interfaces.

Social presence & size & character
relationship from work based site.

$\text{ministry}$ effect

Teacher presence.

Expl. t/l
Coy Sci IL.

Ex - para for kindness
beginning

- Feeding
- Puff n stuff
- Red Flag
- LJP

Thursday, 2 September 2010
Meeting capture

Room where the subjects meet. These microphones are connected to preamplifying equipment in the observation room. This prepares the signal for sound recording either on wire or on plastic discs. (Wire is used when high fidelity and temporary storage is desired; the plastic discs are used when lower fidelity is acceptable and ease of handling and storage is more important.) The signal is also fed into a power amplifier and to a monitoring speaker for the observers. The volume does not have to be kept low since both rooms are sound treated and the connecting mirrors are backed with a second clear pane to further insulate the rooms from each other. The observers therefore confer with each other in low tones without being heard in the experimental room. This is a very great advantage in training and complex observation, since as many as six observers can be accommodated for different kinds of tasks. A team of observers of this size could hardly operate in the same room with the subjects. In general, two interaction observers are used in order to check reliability. Two Interaction Recorders are available. Usually numbers on holders are placed in front of the subjects, for the convenience of the interaction observer who uses the numbers to identify the subjects.

Illustration 3, page 4, is a picture of the Interaction Recorder. This apparatus consists essentially of a case containing a driving mechanism for a wide paper tape upon which scores can be written. A detachable glass plate containing the list of categories fits on the top of the case. At the right side of the list, in proper position for marking down scores or checks, the moving tape is exposed. As each score is put down, it moves with the tape under the check list and disappears, leaving the entire space following the list of categories free for writing again. Inside the case, a marker puts an inked mark across the tape at the end of each minute, and a counter on the switch panel shows the number of minutes that have been recorded. The panel also contains a small ruby light which flashes momentarily once a minute as a signal to the observer to canvas the group for expressive tension behavior which he might otherwise miss in follow-through. Mechanical details of the Interaction Recorder and directions for construction have previously been published (5).
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
• 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)
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

Thursday, 2 September 2010
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

Transcription

Dialogue act

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

Addressee: All
Reflexivity:

Dialogue act

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

Addressee: All
Reflexivity:

Adjacency Pairs

Source: Everybody ready
Type: Request Support
POS: Inform
Target: I think so

Adjacency Pairs

Source: Everybody ready
Type: Request Support
POS: Inform
Target: I think so
Nods

Thursday, 2 September 2010
Gesturing while speaking
Recognition

“A meeting is an event where minutes are taken and hours wasted.”
Multimodal recognition

- Diarization
- Multi-camera tracking
- Activity discovery
- Head pose and visual focus of attention
- Multi-view face detection and recognition
- Gesture and action recognition
Meeting scenarios

- Moving participants
- Rooms from small to large
- Multiple concurrent speakers
Rooms

- No standard
- Diversity in instrumentation
- Acoustic conditions
  - Reverberation
  - Noise
- Size

For audio data capture, all microphones not belonging to the NIST Mark III are connected to a number of RME Octamic eight-channel pre-amplifiers/digitizers. The pre-amplifier outputs are sampled at 44.1 kHz and 24 bits per sample, and are recorded to a computer in WAV format via an RME Hammerfall HDSP9652 I/O card. The 64-channel NIST Mark III data are similarly sampled and recorded in SPHERE format, but are fed into a recording computer via an ethernet connection in the form of multiplexed IP packets.

2.1.2 Video sensor setup

The video data is captured by five fixed cameras. Four of them are mounted close to the corners of the room, by the ceiling, with significantly overlapping and wide-angle fields-of-view. These are set in such a fashion, so that any person in the room is always visible. The fifth camera is mounted on the ceiling and is a Panoramic Camera (PTZ) to capture the entire room from above.

Fig. 2 Schematic diagram of the IBM smart room, one of the five installations used for recording the CHIL corpus. The room is approximately 9 m$^3$ in size and contains nine cameras and 152 microphones for data collection.
Speech in meetings

• Speech detection
• Speaker Identification
• Diarization
• Localization
• Emotion recognition
• and of course speech recognition

“People who enjoy meetings should not be in charge of anything.” (Thomas Sowell)
Person identification

- Obviously multimodal
- Acoustic/Video based
  - A function of the population BUT in practice (for conference room meetings) population is small
  - Advantage: Large amounts of data
- NIST CLEAR evaluations (2006 - )
- Acoustic Person Identification
NIST CLEAR 2007
Acoustic Person ID - Results

Training set 28 individuals, 30s

Stiefelhagen et al. The CLEAR 2007 Evaluation

Thursday, 2 September 2010
Localisation

• NIST CLEAR 2007 evaluations
• Acoustic 3D person tracking

NIST CLEAR 2007

Person tracking - Results

- MOT multiple object tracking metrics
  - precision (MOTP) average distance
  - accuracy (MOTA) assignment based segmental error

Stiefelhagen et al. The CLEAR 2007 Evaluation
Diarization

Who spoke when

• This tasks combines both speaker clustering and segmentation (NOT Speaker identification)

• Only sensible for distant microphones
  • Crosstalk on close talking mic’s not strong enough for speaker confusion
  • Otherwise degeneration into segmentation task (VAD)

• Assessment Diarization Error Rate (DER)
  • Frame error based metric that counts segments as wrong if assigned to the wrong speaker
  • Tries to find best match between ref and hypothesis

• Very challenging task, DER values range from 10 to 30!
Diarization

Most common approach

- Far field pre-processing
  - Beamforming similar or identical to ASR
  - However, keep both enhanced audio and secondary information about position ("delay features")
- Raw segmentation using standard and position features
- Cluster and merge (e.g. using BIC)
- Smoothing
The AMIDA 2009 system
(Huijbregts and van Leeuwen, 2009)
Diarization
AMIDA system performance

- RT 2009 Evaluation test set
- Results
  - Single Distant Microphone (SDM) 29.0 %DER
  - Multiple Distant Microphone (MDM) 21.5 %DER
  - MDM, only main meeting room: 23.2 %DER
  - MDM, no delay features: 28.8 %DER

Thursday, 2 September 2010
Emotion recognition

The AMI experience

Theory

Practice

Emotions are very difficult to annotate and thus it's not clear what to recognise!
Emotion annotation

Feeltrace

• As suggested by the research network humaine.

• Primary signals are detectable though (e.g. laughter)
Speech recognition

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

hmm

rustle

click
“ASR Complete” problem

• Transcription of conversational speech
• Rich in non-native speakers
• Distant speech recognition with microphone arrays
• Speech separation, multiple acoustic channels
• Reverberation
• Overlap detection
• Utterance and speaker segmentation
• Disfluency detection
Meeting data

- Resources by now in several languages
  - e.g. Japanese (CSJ)
  - But we concentrated on English where more than 200 hours are available

- Main corpora

<table>
<thead>
<tr>
<th>Name</th>
<th>Speech (hrs)</th>
<th>User microphones</th>
<th>Distant microphones</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSI</td>
<td>~70</td>
<td>head mounted</td>
<td>4, spaced far apart</td>
<td>LDC</td>
</tr>
<tr>
<td>NIST (2 parts)</td>
<td>~25</td>
<td>head mounted</td>
<td>several arrays, table-top</td>
<td>LDC</td>
</tr>
<tr>
<td>ISL</td>
<td>~11</td>
<td>Lapel</td>
<td>varies, mostly 2</td>
<td>LDC</td>
</tr>
<tr>
<td>AMI</td>
<td>~100</td>
<td>Lapel + head mounted</td>
<td>two 8-microphone arrays</td>
<td>Free</td>
</tr>
<tr>
<td>AMIDA</td>
<td>~7</td>
<td>Lapel + head mounted</td>
<td>two 8-microphone arrays</td>
<td>Free</td>
</tr>
<tr>
<td>CHIL</td>
<td>? (&gt;30)</td>
<td>head mounted</td>
<td>several arrays, table-top</td>
<td>ELRA</td>
</tr>
</tbody>
</table>

- Test sets from LDC, Virginia Tech
Some data properties

<table>
<thead>
<tr>
<th>Meeting resource</th>
<th>Avg Dur (sec)</th>
<th>Avg. Words/Seg</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSI</td>
<td>2.11</td>
<td>7.30</td>
</tr>
<tr>
<td>NIST</td>
<td>2.26</td>
<td>7.17</td>
</tr>
<tr>
<td>ISL</td>
<td>2.36</td>
<td>8.77</td>
</tr>
<tr>
<td>AMI</td>
<td>3.29</td>
<td>10.09</td>
</tr>
<tr>
<td>VT</td>
<td>2.49</td>
<td>8.27</td>
</tr>
<tr>
<td>CHIL</td>
<td>1.80</td>
<td>5.63</td>
</tr>
</tbody>
</table>

- Speaking rate between for all corpora, varying between 3.1 and 3.6 words per second
- 10-15% of data is overlapped speech
  - depending on counting
AMI corpus

- ASR relevant
  - High quality annotation
  - High quality recording
  - sample synchronous on all acoustic channels
  - Balanced test set from 3 locations
  - Scenario vs non-Scenario
  - Contrast
  - Largest meeting corpus available
  - Metadata

- Not the most ‘challenging’
  - Naturalness
Vocabulary

• Differences between corpora
• OOV rates of vocabulary source on domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Vocabulary Source</th>
<th>ICSI</th>
<th>NIST</th>
<th>ISL</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSI</td>
<td>ICSI</td>
<td>0.00</td>
<td>4.95</td>
<td>7.11</td>
<td>6.83</td>
</tr>
<tr>
<td>NIST</td>
<td>NIST</td>
<td>4.50</td>
<td>0.00</td>
<td>6.50</td>
<td>6.88</td>
</tr>
<tr>
<td>ISL</td>
<td>ISL</td>
<td>5.12</td>
<td>5.92</td>
<td>0.00</td>
<td>6.68</td>
</tr>
<tr>
<td>AMI</td>
<td>AMI</td>
<td>4.47</td>
<td>4.39</td>
<td>5.41</td>
<td>0.00</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL</td>
<td>1.60</td>
<td>4.35</td>
<td>6.15</td>
<td>5.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain</th>
<th>Vocabulary Source</th>
<th>ICSI</th>
<th>NIST</th>
<th>ISL</th>
<th>AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSI</td>
<td>ICSI</td>
<td>0.01</td>
<td>0.47</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>NIST</td>
<td>NIST</td>
<td>0.43</td>
<td>0.09</td>
<td>0.59</td>
<td>0.66</td>
</tr>
<tr>
<td>ISL</td>
<td>ISL</td>
<td>0.41</td>
<td>0.37</td>
<td>0.03</td>
<td>0.57</td>
</tr>
<tr>
<td>AMI</td>
<td>AMI</td>
<td>0.53</td>
<td>0.53</td>
<td>0.58</td>
<td>0.30</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL</td>
<td>0.16</td>
<td>0.42</td>
<td>0.53</td>
<td>0.55</td>
</tr>
</tbody>
</table>

source words only

• Vocabulary Padding
• Augment source words with most frequent words from broadcast news sources (up to 50k words)
Content differences

- Language model perplexities
- Meeting resource specific vocabularies
- Interpolation with very basic BN LM

<table>
<thead>
<tr>
<th>Domain</th>
<th>ICSI</th>
<th>NIST</th>
<th>ISL</th>
<th>AMI</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICSI</td>
<td>68.2</td>
<td>74.5</td>
<td>73.7</td>
<td>77.1</td>
<td>67.9</td>
</tr>
<tr>
<td>NIST</td>
<td>105.9</td>
<td>100.9</td>
<td>102.0</td>
<td>105.9</td>
<td>101.2</td>
</tr>
<tr>
<td>iSL</td>
<td>104.7</td>
<td>99.4</td>
<td>98.4</td>
<td>106.3</td>
<td>102.8</td>
</tr>
<tr>
<td>AMI</td>
<td>115.6</td>
<td>114.2</td>
<td>114.4</td>
<td>88.9</td>
<td>94.0</td>
</tr>
<tr>
<td>LDC</td>
<td>97.8</td>
<td>90.7</td>
<td>88.8</td>
<td>92.4</td>
<td>93.8</td>
</tr>
<tr>
<td>ALL</td>
<td>107.5</td>
<td>105.9</td>
<td>105.7</td>
<td>90.6</td>
<td>92.7</td>
</tr>
</tbody>
</table>
AMI corpus

Movement annotation

- Sitting and standing
- Head movement (yes/no)

<table>
<thead>
<tr>
<th></th>
<th>sit</th>
<th>stand</th>
</tr>
</thead>
<tbody>
<tr>
<td>still</td>
<td>74%</td>
<td>7%</td>
</tr>
<tr>
<td>move</td>
<td>15%</td>
<td>4%</td>
</tr>
</tbody>
</table>

- Experiments with 6 hour test sets
- representative or equal proportions
- with and without overlapped segments
AMI corpus

The microphones

<table>
<thead>
<tr>
<th>Audio source</th>
<th># mic</th>
<th>equal</th>
<th>repr</th>
</tr>
</thead>
<tbody>
<tr>
<td>close</td>
<td>1</td>
<td>26.8</td>
<td>30.0</td>
</tr>
<tr>
<td>distant</td>
<td>1</td>
<td>60.2</td>
<td>62.9</td>
</tr>
<tr>
<td>beamforming</td>
<td>2</td>
<td>54.6</td>
<td>57.0</td>
</tr>
<tr>
<td>beamforming</td>
<td>4</td>
<td>52.5</td>
<td>54.3</td>
</tr>
<tr>
<td>beamforming</td>
<td>8</td>
<td>50.8</td>
<td>51.8</td>
</tr>
</tbody>
</table>

AMI corpus, 6 hour test set, no overlap

- Baseline experiments
  - Random selection of test set according to target split, rest is training set
  - ML training / 2009 language models
Influence of overlap

- Training and test set selection with and without segment overlap
- Scoring only non-overlap regions in both cases!
Sit still and wait!

<table>
<thead>
<tr>
<th>Audio source</th>
<th># mic</th>
<th>sit still</th>
<th>stand still</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>no ovl</td>
<td>ovl</td>
</tr>
<tr>
<td>close</td>
<td>1</td>
<td>24.9</td>
<td>35.1</td>
</tr>
<tr>
<td>distant</td>
<td>1</td>
<td>55.9</td>
<td>64.0</td>
</tr>
<tr>
<td>beamforming</td>
<td>2</td>
<td>50.3</td>
<td>59.3</td>
</tr>
<tr>
<td>beamforming</td>
<td>4</td>
<td>47.6</td>
<td>58.4</td>
</tr>
<tr>
<td>beamforming</td>
<td>8</td>
<td>43.5</td>
<td>55.4</td>
</tr>
<tr>
<td>AMI corpus, 6 hour test, equal split</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Overlap and distance are the main sources of degradation
- Head movement has modest influence
AMI ASR systems

- Team from Sheffield, Idiap, Brno, and Edinburgh
  - Active over 5 years
- Acoustic preprocessing and enhancement depends on mic conditions
- Multipass
  - HMM/GMM Acoustic model
  - n-gram language model
- No magic bullet for high accuracy ... more like an acronym shotgun
Front-end

• Task
  • Preprocess audio for ASR

• Standard strategy
  1. produce cleaned up audio - as if recorded with perfect microphone placement
  2. Extract pure speech segments
  3. Attach speaker information for adaptation

Thursday, 2 September 2010
Frontend - Categorisation

- Separation into IHM/MDM/SDM
  - Strategy for most systems: perform speech enhancement on far field channels to yield

- Independent head microphone (IHM)
  - Noise cancelling directional, omnidirectional, lapel
  - Main task - segmentation

- Multiple distant microphones (MDM)
  - Arbitrary number in arbitrary configuration and arbitrary distance
  - Ranges from 2 to 16 in NIST evaluations

- Single distant microphones (SDM)
IHM - Segmentation

- **Signal enhancement (cross-talk suppression)**
- LMS echo cancellation tried but not effective
- Speech activity detection (SAD)

**Using Multi-Layer Perceptron (MLP)**

\[
\begin{align*}
x & \quad \text{(CTM channel)} \\
y_k & \quad \text{(remaining CTM channels)} \\
x' & \quad \text{(enhanced signal)} \\
x' & \quad \text{(36 dim feature vector)} \\
\text{MLP classification} & \\
\text{Viterbi decoder} & \\
\end{align*}
\]
MLP based segmentation

- One of the hardest tasks in meeting recognition!
- MLP
  - Use cross talk energy features
  - Increasing training data to > 100 hours helps!
  - Tuning of hyper-parameters (min state duration, segment insertion penalty, silence collar) by matching with duration histograms for ground truth segmentations
- see Dines et al. (Interspeech 2006)
**Performance**

- 30h / 90h of training data to learn silence!
- 2 pass adapted system

<table>
<thead>
<tr>
<th></th>
<th>#Seg</th>
<th>Tot</th>
<th>CMU</th>
<th>EDI</th>
<th>NIST</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>4527</td>
<td>29.3</td>
<td>36.7</td>
<td>24.5</td>
<td>24.5</td>
<td>31.2</td>
</tr>
<tr>
<td>30h</td>
<td>2717</td>
<td>32.6</td>
<td>41.2</td>
<td>26.2</td>
<td>29.1</td>
<td>33.3</td>
</tr>
<tr>
<td>90h</td>
<td>4541</td>
<td>31.7</td>
<td>42.4</td>
<td>25.3</td>
<td>26.8</td>
<td>31.7</td>
</tr>
</tbody>
</table>

%WER on *rt07seval*.
Performance

- 30h / 90h of training data to learn silence!
- 2 pass adapted system

<table>
<thead>
<tr>
<th>System</th>
<th>$P(sil)$</th>
<th>#Segs</th>
<th>WER %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tot</td>
</tr>
<tr>
<td>ref</td>
<td>–</td>
<td>5660</td>
<td>32.9</td>
</tr>
<tr>
<td>auto</td>
<td>0.80</td>
<td>4809</td>
<td>36.4</td>
</tr>
<tr>
<td>auto</td>
<td>0.85</td>
<td>4949</td>
<td>36.1</td>
</tr>
<tr>
<td>auto</td>
<td>0.90</td>
<td>5135</td>
<td>36.4</td>
</tr>
<tr>
<td>auto</td>
<td>0.95</td>
<td>5504</td>
<td>39.8</td>
</tr>
</tbody>
</table>

%WER on rt09seval.

- and then there was the evaluation ....
Distant microphones

- Enhancement based approach
- Improve audio signal
- Rest identical to close-talking sources
- Optimal microphone configuration not known
  - Beam varies with frequency (and of course geometry)
  - Delay-and-sum based beam-forming most commonly used
- Multi-speaker tracking would allow handling of overlap
  - Complex and errorful
  - Not shown to benefit enough yet
Distant microphones

1. **Gain calibration:** based on peak energy
2. **Noise filtering:** per channel
   - Noise estimate $\theta_{nn}$ based on 20 minimum energy frames
   - Wiener filtering: $H(f) = \frac{\theta_{xx}(f) - \theta_{nn}(f)}{\theta_{xx}(f)}$
3. **Delay estimation:**
   - Scale factor $\alpha_i$ estimation by energy ratio of channel $i$ to reference channel.
   - Delay $\tau_i$ estimation by peak picking in generalised cross correlation
4. **Post-filtering:** Frequency based or heuristics
5. **Beam-forming:** Frame based frequency domain filtering
   \[
   S(f) = \sum_i \alpha_i e^{-2\pi f \tau_i} S_i(f)
   \]
Effect of beamforming

- Unadapted
- Retraining of models mandatory!

<table>
<thead>
<tr>
<th>Audio source</th>
<th># mic</th>
<th>equal</th>
<th>repr</th>
</tr>
</thead>
<tbody>
<tr>
<td>close</td>
<td>1</td>
<td>26.8</td>
<td>30.0</td>
</tr>
<tr>
<td>distant</td>
<td>1</td>
<td>60.2</td>
<td>62.9</td>
</tr>
<tr>
<td>beamforming</td>
<td>2</td>
<td>54.6</td>
<td>57.0</td>
</tr>
<tr>
<td>beamforming</td>
<td>4</td>
<td>52.5</td>
<td>54.3</td>
</tr>
<tr>
<td>beamforming</td>
<td>8</td>
<td>50.8</td>
<td>51.8</td>
</tr>
</tbody>
</table>

AMI corpus, 6 hour test set, no overlap

Thursday, 2 September 2010
Relationship SAD/DER

- SAD has been part of NIST evaluations
- metrics have no penalty on fragmentation
- Relevance for ASR questionable

### Results ICSI 2007

<table>
<thead>
<tr>
<th></th>
<th>SAD (%FER)</th>
<th>Diarisation (%DER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single microphone</td>
<td>14.9</td>
<td>21.7</td>
</tr>
<tr>
<td>Multiple microphones</td>
<td>2.5</td>
<td>8.5</td>
</tr>
</tbody>
</table>

RT’07 best system results
### Relationship DER / WER

<table>
<thead>
<tr>
<th></th>
<th>#clusters</th>
<th>WER (%)</th>
<th>DER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimise for DER</td>
<td>-</td>
<td>60.1</td>
<td>18.1</td>
</tr>
<tr>
<td>Fixed # clusters</td>
<td>6</td>
<td>56.2</td>
<td>30.9</td>
</tr>
<tr>
<td>Fixed # clusters</td>
<td>5</td>
<td>56.1</td>
<td>30.1</td>
</tr>
<tr>
<td>Fixed # clusters</td>
<td>4</td>
<td>55.6</td>
<td>33.6</td>
</tr>
<tr>
<td>Fixed # clusters</td>
<td>3</td>
<td>56.3</td>
<td>38.9</td>
</tr>
<tr>
<td>Fixed # clusters</td>
<td>1</td>
<td>56.9</td>
<td>64.0</td>
</tr>
</tbody>
</table>

- **Poor correlation**
  - Getting the number of speaker clusters roughly right is more important.
## Segmentation performance

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Clustering</th>
<th>Unadapted</th>
<th>Adapted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>-</td>
<td>42.1</td>
<td>36.3</td>
</tr>
<tr>
<td>auto</td>
<td>-</td>
<td>43.8</td>
<td>38.1</td>
</tr>
<tr>
<td>Ref</td>
<td>Ref</td>
<td>40.1</td>
<td>31.1</td>
</tr>
<tr>
<td>auto</td>
<td>no delay</td>
<td>42.8</td>
<td>34.5</td>
</tr>
<tr>
<td>auto</td>
<td>with delay</td>
<td>42.1</td>
<td>32.7</td>
</tr>
</tbody>
</table>

Results AMIDA 2009 on *rt07seval*

- **Standard MLLR adaptation**
- **Modest degradation on well behaved data (**rt07seval**)**
Meeting room attribution

- Remote meeting rooms
  - Sound in remote room is played through speaker
  - Would be recognised twice!

- Solution based on delay in transmission of audio over video conference system
  - Take beam-formed audio file for each room
  - Perform speaker segmentation on room 1 audio

- Estimation of transmission delay possible in practical systems
Meeting room attribution

- For each speaker, for each frame, calculate the max of the cross correlation between the audio from room 1 and room 2 (i.e. the delay).
  - If delay > 0
    - increment room 1 count
  - If delay < 0
    - increment room 2 count
- Assign speaker to room with highest count
Meeting room attribution

Results

<table>
<thead>
<tr>
<th>Description</th>
<th>Segmentation</th>
<th>Tot</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>room-assignment</td>
<td>Auto</td>
<td>33.2</td>
<td>20.6</td>
<td>9.3</td>
<td>3.2</td>
</tr>
<tr>
<td>only room1</td>
<td></td>
<td>36.3</td>
<td>20.5</td>
<td>12.7</td>
<td>3.1</td>
</tr>
<tr>
<td>only room2</td>
<td></td>
<td>45.1</td>
<td>25.8</td>
<td>14.8</td>
<td>4.4</td>
</tr>
<tr>
<td>ref. room-assignment</td>
<td>Ref</td>
<td>30.8</td>
<td>20.1</td>
<td>8.6</td>
<td>2.2</td>
</tr>
<tr>
<td>only room1</td>
<td></td>
<td>33.1</td>
<td>21.0</td>
<td>9.9</td>
<td>2.1</td>
</tr>
<tr>
<td>only room2</td>
<td></td>
<td>41.0</td>
<td>24.3</td>
<td>14.6</td>
<td>2.1</td>
</tr>
</tbody>
</table>

- RT 2009 evaluation test set - rt09seval
- 3 speakers in room 1 - One in room 2

Thursday, 2 September 2010
Language modelling

• Are meetings a domain?
  • Clearly conversational data like CTS
  • Topic range is vast and can be highly specific

• Examples
  • ICSI - “...you already have the super-structure with Gaussians...”
  • NIST - “...well there was the dirty diaper game ...”
  • AMI - “...where you think a remote control could go ...”
  • VT - “...looking at the weather data for Sri Lanka  ...”
  • ISL - “...well , that speaks for the air strike...”
**Challenge Language**

- Meeting source specific LMs seem to make a difference in perplexity but not in WER

<table>
<thead>
<tr>
<th>LM Data</th>
<th>Overall</th>
<th>male</th>
<th>female</th>
<th>Scenario</th>
<th>Non-Scen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast News</td>
<td>99.8</td>
<td>99.3</td>
<td>100.9</td>
<td>87.9</td>
<td>137.8</td>
</tr>
<tr>
<td>CTS</td>
<td>100.5</td>
<td>100.1</td>
<td>101.6</td>
<td>88.2</td>
<td>140.2</td>
</tr>
<tr>
<td>Meetings</td>
<td>102.7</td>
<td>101.6</td>
<td>105.4</td>
<td>91.2</td>
<td>138.8</td>
</tr>
<tr>
<td>Combined (inc Web-Data)</td>
<td>92.9</td>
<td>92.8</td>
<td>93.2</td>
<td>84.1</td>
<td>119.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language model</th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>OtherEU</th>
<th>S. Asia</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast News</td>
<td>105.2</td>
<td>97.7</td>
<td>128.5</td>
<td>113.3</td>
<td>112.0</td>
<td>102.8</td>
</tr>
<tr>
<td>CTS</td>
<td>105.9</td>
<td>100.2</td>
<td>128.9</td>
<td>114.4</td>
<td>115.0</td>
<td>104.0</td>
</tr>
<tr>
<td>Meetings</td>
<td>110.3</td>
<td>98.0</td>
<td>126.8</td>
<td>115.9</td>
<td>113.3</td>
<td>103.7</td>
</tr>
<tr>
<td>Combined</td>
<td>96.9</td>
<td>90.8</td>
<td>111.0</td>
<td>103.0</td>
<td>104.7</td>
<td>94.9</td>
</tr>
</tbody>
</table>
### Challenge Data

<table>
<thead>
<tr>
<th>LM components</th>
<th>#words</th>
<th>Meetings</th>
<th></th>
<th>Lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI data (prelim)</td>
<td>206K</td>
<td>0.051</td>
<td>0.038</td>
<td>0.040</td>
</tr>
<tr>
<td>CHIL</td>
<td>76K</td>
<td></td>
<td></td>
<td>0.167</td>
</tr>
<tr>
<td>Fisher</td>
<td>21M</td>
<td>0.214</td>
<td>0.237</td>
<td>0.219</td>
</tr>
<tr>
<td>Hub4 LM96</td>
<td>151M</td>
<td>0.028</td>
<td>0.044</td>
<td>0.051</td>
</tr>
<tr>
<td>ICSI meeting corpus</td>
<td>0.9M</td>
<td>0.093</td>
<td>0.080</td>
<td>0.067</td>
</tr>
<tr>
<td>ISL meeting corpus</td>
<td>119K</td>
<td>0.126</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>NIST meeting corpus</td>
<td>157K</td>
<td>0.085</td>
<td>0.065</td>
<td>0.064</td>
</tr>
<tr>
<td>Switchboard/Callhome</td>
<td>3.4M</td>
<td>0.057</td>
<td>0.070</td>
<td>0.063</td>
</tr>
<tr>
<td>webdata (meetings)</td>
<td>128M</td>
<td>0.198</td>
<td>0.163</td>
<td>0.155</td>
</tr>
<tr>
<td>webdata (fisher)</td>
<td>128M</td>
<td>0.066</td>
<td>0.103</td>
<td>0.144</td>
</tr>
<tr>
<td>webdata (rt06s-conf)</td>
<td>138M</td>
<td>0.081</td>
<td>0.108</td>
<td>0.106</td>
</tr>
<tr>
<td>webdata (rt06s-lect)</td>
<td>114M</td>
<td></td>
<td></td>
<td>0.036</td>
</tr>
<tr>
<td>rto5seval perplexity</td>
<td></td>
<td>105.6</td>
<td>84.3</td>
<td>81.2</td>
</tr>
</tbody>
</table>

- LM interpolation weights optimised on conference and lecture meetings
LM - Conclusions

- Web data is very important
  - Developed complimentary retrieval (Wan and Hain, 2006)
- Reality is cruel!
  - Development result using LMs from years 2005-2007. Each year the amount of data was doubled and the perplexity decreased by ~10%.

<table>
<thead>
<tr>
<th>rt07seval</th>
<th>TOT</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>CMU</th>
<th>EDI</th>
<th>NIST</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>lm05</td>
<td>28.7</td>
<td>14.9</td>
<td>10.2</td>
<td>3.5</td>
<td>33.6</td>
<td>20.7</td>
<td>14.7</td>
<td>31.7</td>
</tr>
<tr>
<td>lm06</td>
<td>28.6</td>
<td>14.9</td>
<td>10.2</td>
<td>3.5</td>
<td>34.1</td>
<td>20.2</td>
<td>14.4</td>
<td>31.5</td>
</tr>
<tr>
<td>lm07</td>
<td>28.5</td>
<td>14.8</td>
<td>10.2</td>
<td>3.5</td>
<td>34.0</td>
<td>20.3</td>
<td>14.4</td>
<td>31.1</td>
</tr>
</tbody>
</table>

- Has more to do with rt07seval
Acoustic modelling

The acronym graveyard

Acoustic modelling
Acoustic modelling Fundamentals

• Amounts of data
  • State of the art systems require many hundreds of hours of training data
  • Augment with background corpora

• What to train on
  • Training of IHM/SDM/MDM differently
  • Segmentation usually poor on all corpora

• Specific issues
  • MDM/SDM training: special strategies for overlapped speech
  • Data integrity: Filtering of poor quality segments due to
Where are the gains?

- Basic channel related techniques
  - Cepstral mean and variance normalisation (also for MDM!)
  - Posterior features
  - Adaptation from different domain, different bandwidth data (CTS)
- Speaker specific modelling
  - Vocal tract length normalisation
  - SAT
- Enhanced parameter estimation
  - MPE, fMPE, HLDA

Thursday, 2 September 2010
Cross domain adaptation

- Adaptation of CTS models (300h) to meeting domain (100h)
- Inclusion of HLDA adaptation (Karafiat et al, 2007)

<table>
<thead>
<tr>
<th>Initial models</th>
<th>Adaptation</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS SAT MPE</td>
<td>ML-MAP</td>
<td>26.0</td>
</tr>
<tr>
<td>CTS SAT MPE + ML-MAP</td>
<td>MPE-MAP</td>
<td>23.9</td>
</tr>
</tbody>
</table>

Results on rt05seval
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

- Variants with Bottle-neck MLP (LCRCBN)
- Stacked bottleneck (SBN) uses one MLP for translation of window and one for the merge.
Results (RT07, IHM)

- ML
- MPE
- fMPE
- fMPE + MPE

Training

WER/%

- HLDA-PLP
- HLDA-PLP + LCRC

Thursday, 2 September 2010
Adaptation

- Standard MLLR / CMLLR adaptation
- Slightly reduced gain with posterior features

<table>
<thead>
<tr>
<th>Features</th>
<th>Tr</th>
<th>Adapt/Normalise</th>
<th>TOT</th>
<th>CMU</th>
<th>EDI</th>
<th>NIST</th>
<th>TNO</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>ML</td>
<td>ML</td>
<td>39.7</td>
<td>39.9</td>
<td>37.0</td>
<td>34.2</td>
<td>38.9</td>
<td>45.8</td>
</tr>
<tr>
<td>MFCC</td>
<td>ML</td>
<td>VTLN HLDA</td>
<td>34.2</td>
<td>34.2</td>
<td>32.6</td>
<td>29.9</td>
<td>32.0</td>
<td>41.0</td>
</tr>
<tr>
<td>MFCC + BN</td>
<td>ML</td>
<td>VTLN HLDA</td>
<td>29.4</td>
<td>29.3</td>
<td>27.5</td>
<td>26.6</td>
<td>28.1</td>
<td>35.6</td>
</tr>
<tr>
<td>MFCC + BN</td>
<td>ML</td>
<td>VTLN HLDA SAT</td>
<td>27.3</td>
<td>27.2</td>
<td>25.2</td>
<td>25.6</td>
<td>26.5</td>
<td>32.3</td>
</tr>
<tr>
<td>MFCC + BN</td>
<td>MPE</td>
<td>VTLN HLDA SAT</td>
<td>25.6</td>
<td>25.6</td>
<td>23.0</td>
<td>23.6</td>
<td>24.9</td>
<td>30.1</td>
</tr>
</tbody>
</table>

- Adaptation to complete meeting brings little gain
  - Intersection of complementary systems helps a bit
Decoding - Juicer

- Juicer is a WFST based decoder, started at IDIAP, now jointly developed by IDIAP, Edinburgh, and Sheffield
  - capable of online decoding
- Juicer requires LM pruning but can use 4g

![Graph showing perplexity on RT09 ihm text]
## Speed and WFST size

Entropy pruning and decoding using fixed beam settings

<table>
<thead>
<tr>
<th>Total n-grams in 4g LM</th>
<th>Arcs in WFST</th>
<th>WER</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5M</td>
<td>15.6M</td>
<td>46.8</td>
<td>0.579</td>
</tr>
<tr>
<td>4.4M</td>
<td>19.5M</td>
<td>46.5</td>
<td>0.591</td>
</tr>
<tr>
<td>6.1M</td>
<td>26.6M</td>
<td>46.6</td>
<td>0.597</td>
</tr>
<tr>
<td>8.0M</td>
<td>35.2M</td>
<td>46.7</td>
<td>0.606</td>
</tr>
</tbody>
</table>

## Change of lexicon size and LM order

<table>
<thead>
<tr>
<th>Lexicon size</th>
<th>LM orde</th>
<th>Arcs in WFST</th>
<th>WER</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2K</td>
<td>7</td>
<td>11.8M</td>
<td>55.3</td>
<td>0.827</td>
</tr>
<tr>
<td>6K</td>
<td>7</td>
<td>12.5M</td>
<td>48.2</td>
<td>0.625</td>
</tr>
<tr>
<td>10K</td>
<td>7</td>
<td>13.8M</td>
<td>47.2</td>
<td>0.582</td>
</tr>
<tr>
<td>16K</td>
<td>7</td>
<td>14.7M</td>
<td>46.8</td>
<td>0.589</td>
</tr>
<tr>
<td>50K</td>
<td>4</td>
<td>15.6M</td>
<td>46.8</td>
<td>0.579</td>
</tr>
</tbody>
</table>

*Build the biggest LM using the minimal OOV for any given speed!*
Decoder optimisation
Automatic optimisation of decoder parameters

Gradient based optimisation of 8 pruning parameters

Optimal configuration for fast decoding
Decoder optimisation

Baseline HDecode w/2g & HLRescore w/4g, LM scale=14, -5
Baseline HDecode w/2g & HLRescore w/4g, LM scale=15,0

Thursday, 2 September 2010
Systems

• Transcription of speech in meetings from a large number of different sites
• Different channels: close-talking headset, lapel, microphone array, single microphone far field.
• Different topics
• Mix of English accents
• Different target metrics (STT/SASTT)
• System designs changed
Example 2005 System

IHM
- Headset microphone recordings
  - Multi-channel echo cancellation
  - MLP based segmentation
  - Smoothing
  - Frontend Processing
    - MLE, tgcomb05
    - Resegmentation
    - VTLN, CMN, CVN
  - MPE triphones, VTLN, SHLDA, tgcomb05

MDM
- Tabletop microphone recordings
  - Delay vector estimation
  - Delay-Sum beamforming
  - Speaker segmentation/clustering
  - Bigram Lattices
  - MRS lattice expansion
  - 4-gram Lattices
  - MLLR, 2 speech transforms
    - MPE triphones, VTLN, SHLDA, Pronprob
    - 4-gram Lattices
    - CN generation
    - Minimum word error rate decoding
    - Alignment
Example 2007 System

- Design paradigm
- Cross adaptation
- System combination
- On the slow side
  - > 100 x RT
- Masssive AM training
  - 2000h CTS
  - 200h Meetings
- Very good IHM results
2009 Realisation - ROTK

Developed module graph based systems

- Organises
- Data flow
- Parametrisation
- Compute distribution
2009 System

- Beamforming (MDM)
- Speech/non-speech segmentation
- Speaker clustering (MDM)
- PLP and SBN features
- HMM/GMM system (122k 39D Gaussians)
  - HLDA, VTLN, SAT, MPE, fMPE
- Meeting data only!
- 50k vocabulary
- Trigram language model
- Weighted FST decoder
- Many passes partially automatically optimised
Results (RT09, MDM)

RT09 evaluation, mic array

Meeting Recording

WER/%

0 5 10 15 20 25 30 35 40 45 50

UEdin-1  UEdin-2  Idiap-1  Idiap-2  NIST-1  NIST-2  NIST-3
Results (RT09, MDM)

RT09 evaluation, mic array
- Including overlapping speech
- Non-overlapped segments only

WER/%

Meeting Recording

UEdin-1  UEdin-2  Idiap-1  Idiap-2  NIST-1  NIST-2  NIST-3
Examples

• Offline meeting ASR
  • Available online!
  • Embedded for search and indexing
  • Aid for transcripts

• Online recognition
  • Content linking
# Dr Vincent Wan's Account

**Uploaded files (individually processed)**

<table>
<thead>
<tr>
<th>Filename</th>
<th>Upload timestamp</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE4-O08002.wav</td>
<td>27 June 2006, 5:10 pm</td>
<td>✔</td>
</tr>
<tr>
<td>1234.wav</td>
<td>26 August 2009, 11:38 pm</td>
<td>✔</td>
</tr>
<tr>
<td>peter rabbit.wav</td>
<td>26 August 2009, 11:39 am</td>
<td>✔</td>
</tr>
<tr>
<td>55.wav</td>
<td>26 August 2009, 11:54 pm</td>
<td>✔</td>
</tr>
</tbody>
</table>

- Full client path: C:/Users/Vinny/Downloads/55.wav
- Size: 32.00 KB
- Format: Mono,PCM,16Khz,16-bit
- Upload status: complete
  - Systems: s0001 (first label system)  
    - Date: 26 August 2009, 11:54 pm
    - Transcripts: submitted.html

**Account summary**
- Total number of files uploaded: 178
- Total size uploaded: 2.33 GB
- Upload limit: 20.00 GB every 30 days
- Used in last 30 days: 181.35 MB
- Available: 19.62 GB

[Return to user management](http://www.webasr.org)
Audio Notetaker

... doing was buying ... it turned out ... somebody said to me if you're in trouble getting cheap flights ... travel agents ... try going to the sides ... really ... quicker than travel agent ... really have ... right ... did ... right ... went ... sides ... think he has ... yes we could just ... right ... ... think ... ... doing what we're doing sitting ... ... work ... kind of happened over series of ... ... actually for ... okay ... ... lack of all the ... right well the first ... work which means that I would never buy anything ... comfortable about ups ... suppose it comes about using ... buy things in cases seems inappropriate to use ... ... know that ... small ... work ... don't know what checks because I saw ... network ... sorts of things ... wants to try ... quick look at things one on
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)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spkr Turn Feats</td>
<td>55.1%</td>
</tr>
<tr>
<td>Lexical Feats</td>
<td>48.7%</td>
</tr>
<tr>
<td>Prosodic Feats</td>
<td>59.6%</td>
</tr>
<tr>
<td>Acoustic Comb</td>
<td>44.2%</td>
</tr>
<tr>
<td>Visual Feats</td>
<td>59%</td>
</tr>
<tr>
<td>MM Comb</td>
<td>43.6%</td>
</tr>
</tbody>
</table>
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>43.6</td>
</tr>
<tr>
<td>Multistream</td>
<td>13.5</td>
</tr>
<tr>
<td>Multistream + duration model</td>
<td>12.2</td>
</tr>
</tbody>
</table>

- 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

Thursday, 2 September 2010
Content linking

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

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

Apple Mouse
Banana
Abstract factory pattern
Wikipedia:Reference desk/Archive
Meeting recording (2010)
Commodity mic arrays
Ambient spotlight
Conclusions

- The AMI corpus is a great resource [http://corpus.amiproject.org](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, Jonathan Kilgour, Mike Lincoln, Andrei Popescu-Belis, Lukas Burget, Martin Karafiat, Asmaa El Hannani, Vincent Wan, Iain McCowan, John Dines, Philip N. Garner

- Some current and former PhD students:
  Alfred Dielmann, Giulia Garau, Songfang Huang, Gabriel Murray, Le Zhang, Erich Zwyssig, Davide Marino, M. Gibson

- Other supporting projects

Thursday, 2 September 2010