The AMI Meeting Transcription System

AMI
AUGMENTED MULTI-PARTY INTERACTION

Thomas Hain
University of Sheffield

Lukas Burget, Martin Karafiat (Univ Tech Brno)
Giulia Garau, Mike Lincoln, Steve Renals (Univ Edinburgh)
John Dines, Darren Moore, Iain McCowan, Jithendra Vepa (IDIAP)
Vincent Wan (Univ Sheffield)
Roeland Oerdelman (Univ Twente)
David van Leeuwen (TNO)

Cambridge, June 2007

Outline

- AMI/AMIDA and ASR
- Meeting data
- System components
  - Acoustic processing
  - Front-ends
    - Posterior based features
  - Acoustic modelling
    - Discriminative training
    - Cross-domain adaptation
  - Language modelling
    - Web-data
- System architecture and results on RT'07S

AMI/AMIDA

- Objective
  - "Round the table"-meeting analysis
  - Low and high level processing
  - Assistive tools
    - for review
    - for online support (distant access)
- EC funded project(s) : Total duration $5\frac{1}{2}$ years

Meeting Transcription

- Basic properties
  - Conversational speech
  - Unknown number of participants
  - Large variation of sound quality
  - Usually multiple recordings available (head-mounted, lapel, table-top, mic-array, ...)
  - People move in the room (sometimes)
- NIST evaluations definitions
  - *Conference room* vs. *Lecture room*
  - Close-talking (*IHM*)
  - Single distant microphone (SDM)
  - **Multiple distant microphones (*MDM*)**
  - ..

Meeting Corpora

Several corpora are now available for ASR purposes (i.e. including manual transcriptions).

Name	Speech (hours)	User microphones	Distant microphones	Availability
ICSI	˜70	head mounted	4, spaced far apart	LDC
NIST (2 parts)	˜25	head mounted	several arrays, table-top	LDC
ISL	˜11	Lapel	varies, mostly 2	LDC
AMI	˜ 100	Lapel + head mounted	two 8-microphone arrays	Free
CHIL	? (>30)	head mounted	several arrays, table-top	ELRA

- ICSI, and some NIST, AMI, CHIL meetings are “natural”.
- About half of the data is only available since last year.
- Some of the above include more detailed information such as video, speaker location, etc.
- Small amounts are available from Virginia Tech (VT) and the LDC.

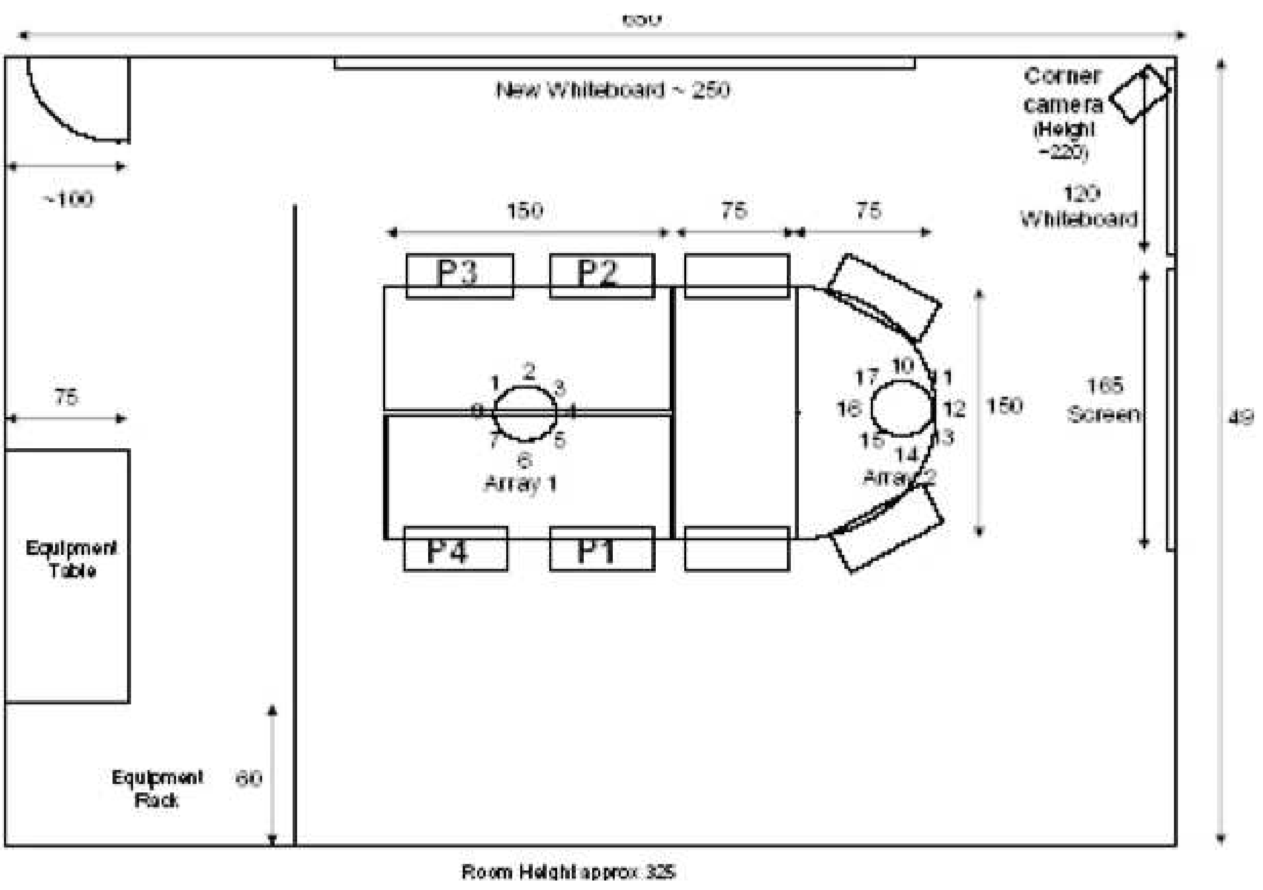
LDC denotes the Linguistic Data Consortium. ELRA stand for European Language Resource Association.

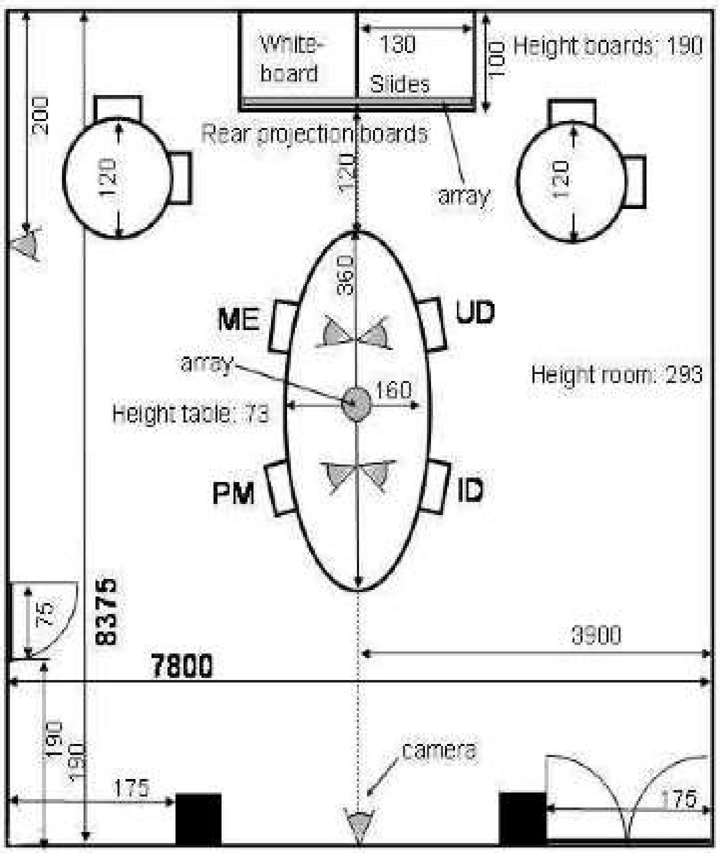
AMI Corpus

- Splits into 70 hour “scenario” and 30 hours “non-scenario” meetings
  - Scenario: 4 consecutive meetings on design of a remote control
- Data:
  - Audio
    - Head-set microphones (low and high quality, noise cancelling)
    - Lapel microphones
    - 2 circular microphone arrays (only one really useful)
    - Manikin
    - Studio quality synchronisation
  - Video, projector information, ...
  - Several layers of annotation (including full text)

http://corpus.amiproject.org

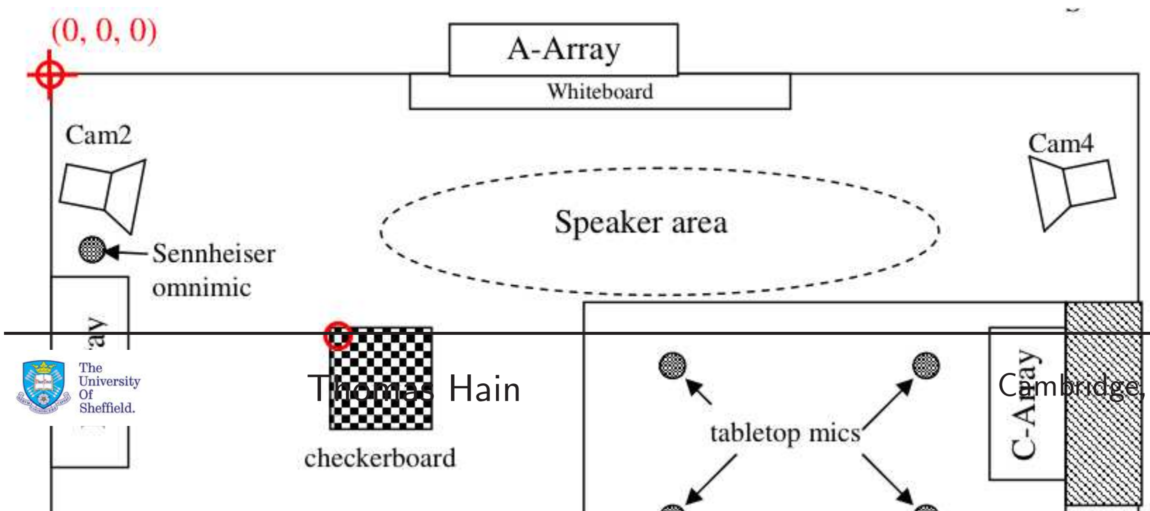
Meeting Rooms - AMI

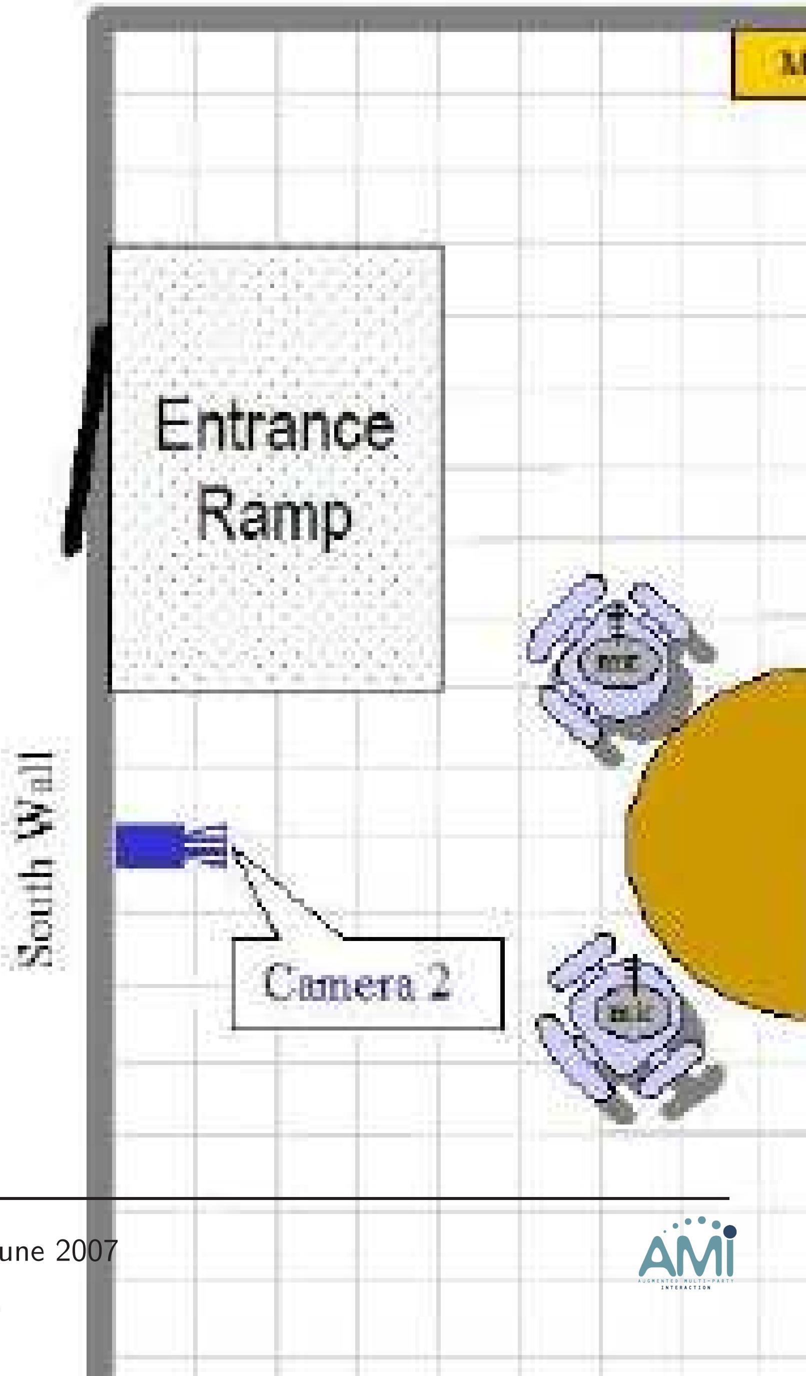




Sample - EN2001b

Other Meeting Rooms


(0, 0, 0)
A-Array
Whiteboard
Cam2
Cam4
Speaker area
Sennheiser
omnimic
ay
checkerboard
tabletop mics
C-Array


Entrance
Ramp
South Wall
Camera 2

- Considerable variation of microphone placement
- Some meeting rooms have only approximate placement information

System Development - Starting Point

- No large vocabulary system
  - Toolkits: HTK 3.3, Brno STK, SRI LM toolkit
  - CUED HDecode (Pre- "public release")
  - CTS training segmentation's/transcripts
  - Data (for acoustic and language modelling)

- Distributed development
  - 5 sites
  - 6 sub-groups

Dictionary and Vocabulary Selection

UNISYN baseline dictionary has pronunciations for 114,876 words (Fitt, 2000)

```
abandon::VB/VBP/NN: { @ .  b * a n .  d @ n } :{abandon}:1986
```

based on the idea of transformation to dialects.

- ≈15k words were added using a combination of automatic and manual generation:

  1. Part-word pronunciations initially automatically guessed from the existing pronunciations
  2. Automatic CART based letter-to-sound conversion trained from UNISYN (Festival)
     8% phone accuracy and 89% word accuracy on a held-out part of the UNISYN dictionary.
     89% phone accuracy and 51% word accuracy on the added data - mostly irregular words.
  3. Hand correction/checking of all automatic hypotheses

- Vocabulary selection by taking all words from meeting domains and padding with most frequent words to obtain 50k dictionary.

http://www.cstr.ed.ac.uk/projects/unisyn

Meeting Domain - Vocabularies

- The content of meetings varies hugely (from games to highly technical meetings)
  - Are Out-Of-Vocabulary (OOV) rates a problem ?

Raw	OOV Rates			
Corpus	ICSI	NIST	ISL	AMI
ICSI	0.00	4.95	7.11	6.83
NIST	4.50	0.00	6.50	6.88
ISL	5.12	5.92	0.00	6.68
AMI	4.47	4.39	5.41	0.00
ALL	1.60	4.35	6.15	5.98

Padding	OOV Rates			
Domain	ICSI	NIST	ISL	AMI
ICSI	0.01	0.47	0.58	0.57
NIST	0.43	0.09	0.59	0.66
ISL	0.41	0.37	0.03	0.57
AMI	0.53	0.53	0.58	0.30
ALL	0.16	0.42	0.53	0.55

- Padding with words from broadcast news levels OOV rates.

Meeting Domain - Language Modelling

- Linearly interpolated N-gram language models
- Meeting corpus specific language models

	Language model Perplexities					
Data source	ICSI	NIST	ISL	AMI	LDC	fgcomb05
ICSI	**82.7**	86.2	87.3	97.1	109.9	84.2
NIST	101.4	**103.7**	102.0	105.7	109.2	98.9
ISL	110.1	111.0	**106.7**	119.3	114.5	108. 6
AMI	92.9	108.9	108.7	**77.3**	101.7	84. 1
LDC	92.4	92.8	87.6	99.0	**84.3**	90. 5
ALL	86.9	93.2	93.7	92.1	106.7	**85.4**

fgcomb05 was used in the RT'05 evaluations.

AMI corpus - OOV and Language Modelling

- The AMI corpus has
  - Scenario and Non-scenario meetings
  - Large amounts of non-native speech

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

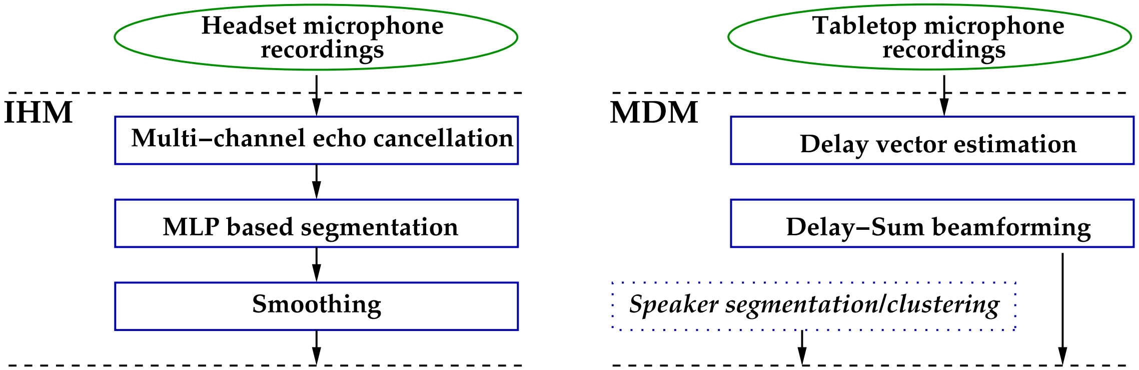
Language model	English	French	German	OtherEU	S. Asia	Rest of World
Broadcast News	105.2	**97.7**	**128.5**	113.3	112.0	102.8
CTS	105.9	**100.2**	**128.9**	114.4	115.0	104.0
Meetings	110.3	**98.0**	**126.8**	115.9	113.3	103.7
Combined (inc Web-Data)	96.9	**90.8**	**111.0**	103.0	104.7	94.9

The above includes part words,without perplexities are usually 10 lower ..

OOV rates are lowest for Germans and highest for French and general EU ...

Front-end Processing

- Tasks
  - Segmentation
  - Speaker clustering (for adaptation)
  - Enhancement

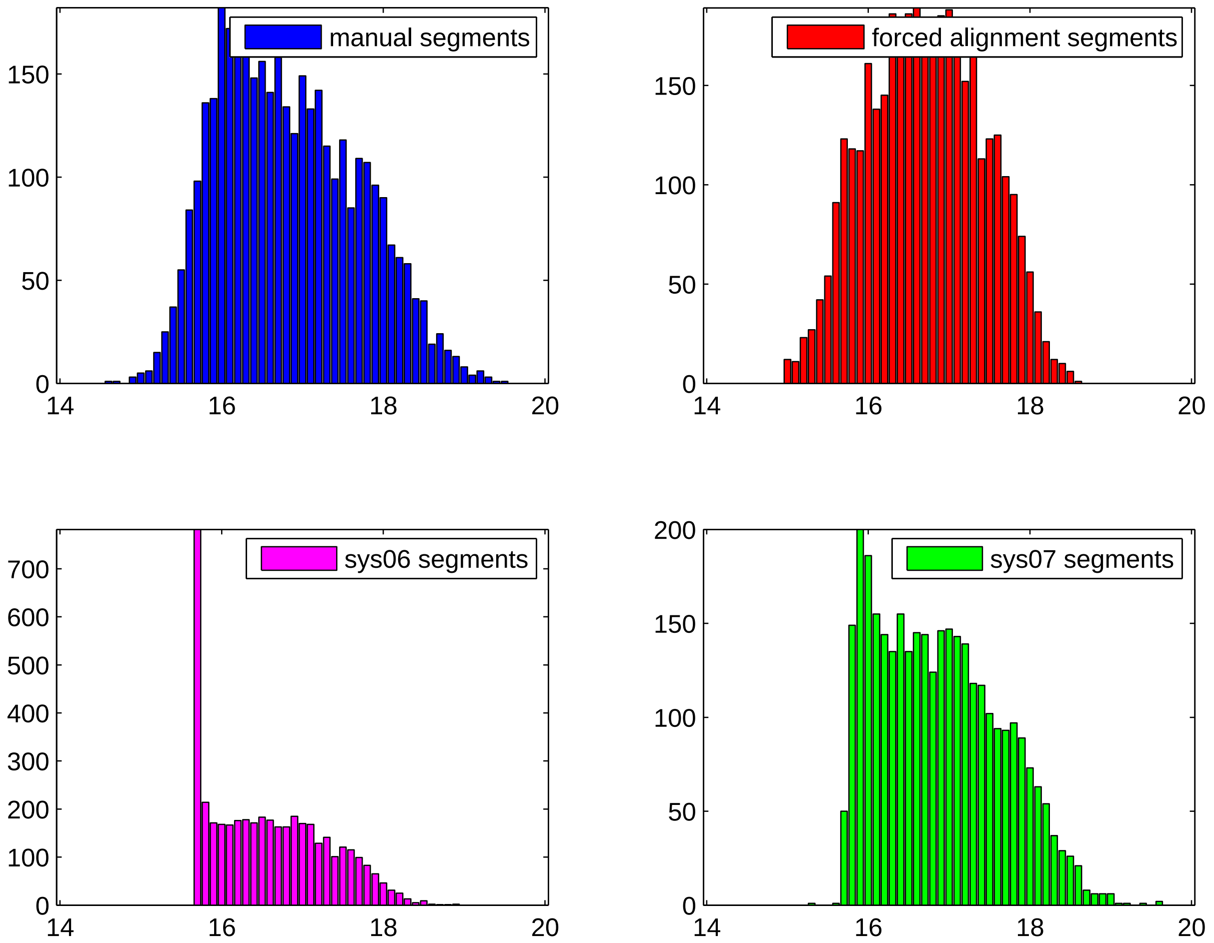

Headset microphone recordings
IHM
Multi−channel echo cancellation
MLP based segmentation
Smoothing
Tabletop microphone recordings
MDM
Delay vector estimation
Delay−Sum beamforming
Speaker segmentation/clustering

- Individual Head Microphones (IHM)
  - huge quality difference between microphones, inc lapel
  - Severe cross-talk
  - Non-speech human noises (breathing)
- Multiple Distant Microphones (MDM)
  - undefined locations of microphones

IHM Processing

- Echo cancellation: Adaptive LMS based signal cross-talk suppression
- **Features**:
  - 13 MF-PLP + energy
  - Cross-channel normalised energy
  - Signal kurtosis
  - Maximum normalised cross-correlation
  - Mean cross-correlation
  - 54D (1st and 2nd order differentials)
- **MLP classifier:**
  - 31 input frames 2 output classes, 50 hidden units
  - 90 hours training, 10 hours cross-validation
- **Segmentation:**
  - Segment minimum duration of 0.5 seconds, added 0.1 second silence collar to segments
- Priors to fit segment histogram

Histogram of Speech Segment Durations



IHM Front-end - RT'06 Performance

- Number of channels per meeting relates to proportion of FA/FR errors

	EDI	TNO	CMU	VIT	NIS	**TOT**
INS	4.1	5.0	6.2	4.7	4.4	**4.9**
DEL	7.7	10.0	8.0	9.0	8.5	**8.5**
SUB	21.1	30.4	29.2	28.0	27.7	**27.0**
WER	32.8	45.4	43.4	41.6	40.6	**40.4**

manual

	TOT
INS	**3.5**
DEL	**9.4**
SUB	**26.5**
WER	**39.3**

re-segmented manual

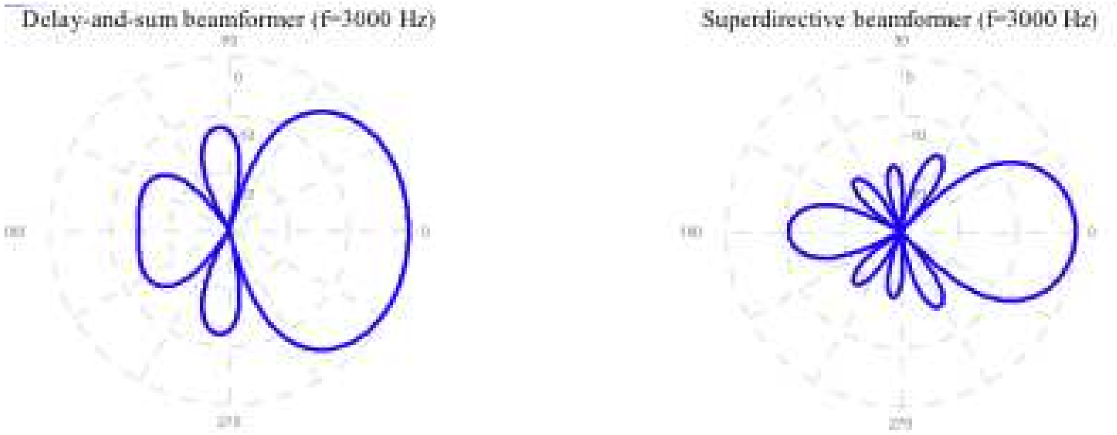
	EDI	TNO	CMU	VIT	NIS	**TOT**
INS	3.8	3.8	4.3	2.7	2.9	**3.5**
DEL	9.6	11.5	10.8	16.1	15.1	**12.6**
SUB	20.3	30.5	28.2	24.8	24.5	**25.3**
WER	33.7	45.9	43.4	43.6	42.5	**41.4**

automatic

Microphone Arrays

Orignal/Beamformed

- Enhancement based approach
  - Improve audio signal
  - Then process 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



Example taken from (I. Tashev. 2006)

- Alternative: Others have used ROVER ....

MDM Processing

- Using multiple microphones for speech enhancement

1. **Gain calibration:** on complete meeting, based on peak energy
2. **Noise filtering**: per channel
   - ▷ Noise estimate θ_{nn} based on 20 minimum energy frames
   - ▷ Wiener filtering: $H(f) = \frac{\theta_{xx}(f) - \theta_{nn}(f)}{\theta_{xx}(f)}$
3. **Delay estimation:**
   - ▷ 1 second frames, 0.5 second frame shift
   - ▷ Scale factor α_i estimation by energy ratio of channel i to reference channel.
   - ▷ Delay τ_i estimation by peak picking in generalised cross correlation
4. **Beam-forming:** Frame based frequency domain filtering

$$S(f) = \sum_i \alpha_i e^{-2\pi f \tau_i} S_i(f)$$

- BUT: In cases of directed microphones or only 2 microphones, simply pick highest energy channel for every time frame

MDM Speaker Segmentation/Clustering

- Segmentation
  - BIC-based voice activity detection on beam-formed channel.
- BIC based clustering

	#clusters	WER (%)	DER (%)
Optimise for DER	-	60.1	**18.1**
Fixed # clusters	6	56.2	30.9
Fixed # clusters	5	56.1	30.1
Fixed # clusters	4	**55.6**	33.6
Fixed # clusters	3	56.3	38.9
Fixed # clusters	1	56.9	64.0

Results on the RT'06 evaluation set

- We confirm that diarisation error rate (DER) and word error rate (WER) are not related.
- A fixed number of clusters yields best results.
- Beamforming data was (not yet) used.

Overlapped Speech

- Concurrent speech is very frequent
  - Causes cross-talk on IHM conditions
  - Distorts for MDM
- Data selection for training of MDM models: Removal of overlapped speech
  - Based on timing from alignments of IHM channels

	#segs	Speech retained (hours)
IHM	238455	172.8
no overlap	-	˜70
WB - 3	191894	134.1
WB - 5	190238	133.2
WB - 10	186625	**131.2**
WB - 20	181890	127.9
WB - 30	177613	124.9

(x in WB x denotes minimum distance from word boundary)

- No system presented yet that targets overlapped speech.

Acoustic Modelling

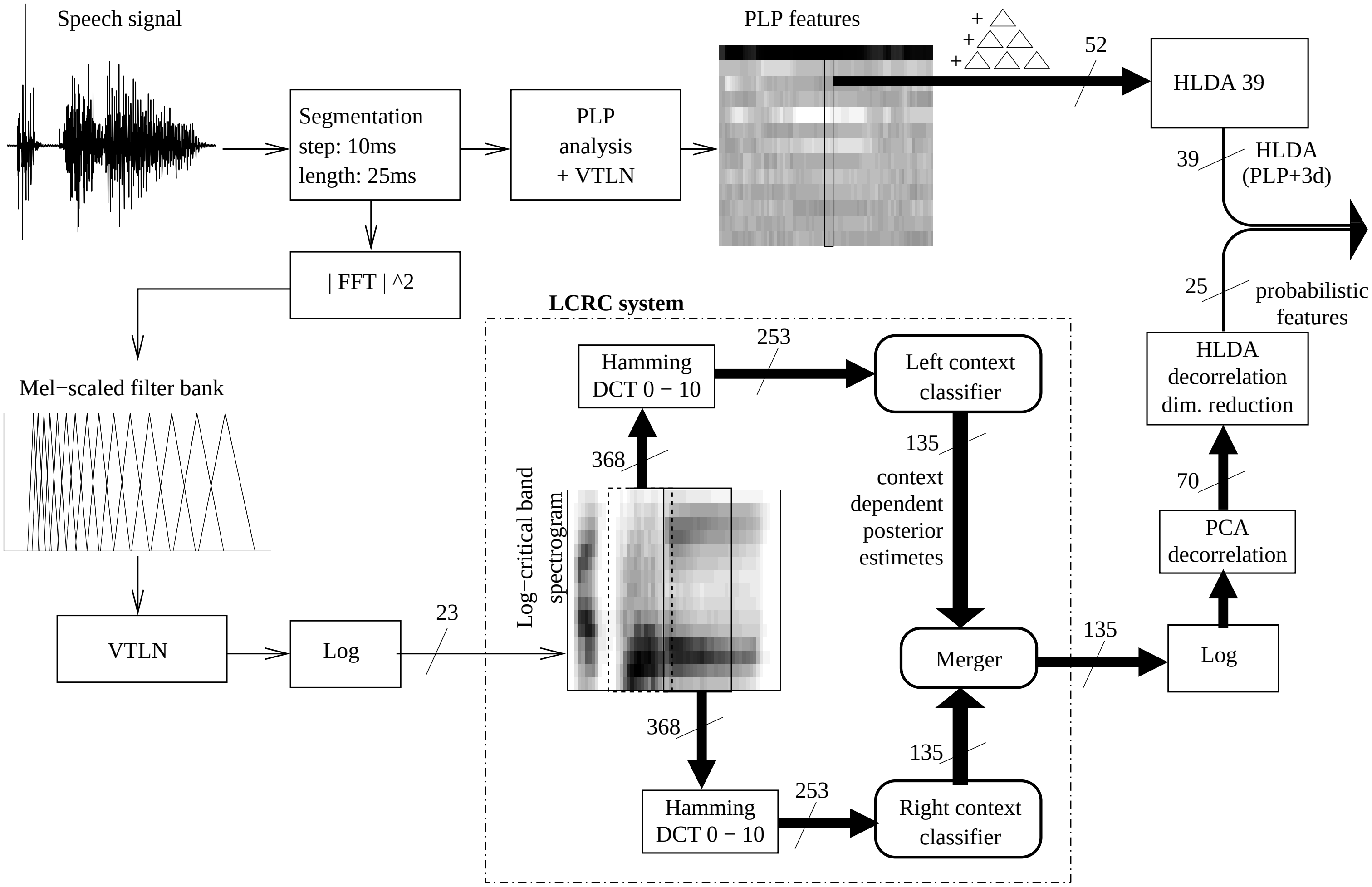
- Standard acoustic modelling techniques are used with similar performance gains to CTS (numbers are relative reductions in word error rate).
  - Heteroscedastic linear discriminant analysis (HLDA) (∼3-5%)
  - Speaker adaptive training (SAT) (∼2%)
  - Vocal tract length normalisation (VTLN) (∼10-15%)
  - MPE training (∼10%)

- Special features in the AMI system
  - Posterior based front-end feature extraction
  - Adaptation from CTS/Fisher models

Discriminative Training - MPE

- Based on STK (Univ. of Technology Brno)
- Features
  - ▷ Model lattice generation with unigram language models (at 2×RT) Language models are trained on training data only.
  - ▷ Acoustic and language model scale factors
    - Scaling of state posteriors !
    - Penalising the language model
  - ▷ Full lattice forward-backward with time based pruning option.
  - ▷ I-smoothing
  - ▷ Merged numerator/denominator lattice
  - ▷ Training iteration operates at 0.2×RT
- Relatively fast generation of lattices allows to rebuild lattices for all tasks

Posterior Features



Posterior Features (2)

- LCRC features (Schwarz,2004)
  - Trained on 100 hours of data
  - MLPs have 1500 hidden units
  - PCA/HLDA combination can be replaced with HLDA using phone level estimation.

- Performance on the RT'05 evaluation set

System	PLP HLDA WER [%]	LC-RC WER [%]
Basic HMM	28.7	25.2
SAT	27.6	23.9
SAT MPE	24.5	21.7

- Alternative: "Bottleneck" features
  - Merging MLP has 5 layers with (1500,30,1500) units fir the hidden layers.
  - Output of the hidden layer is used directly as feature vector.
  - Yields equivalent performance.

Adaptation of CTS Models

- Motivation
  - Smoothing due to substantial increase of training data
- Issues:
  - Narrowband (NB) vs Wideband (WB)
- Solution:
  - Train constrained MLLR (CMLLR) transform from NB to WB data

Data	Bandwidth	Adaptation	#Iter	%WER
CTS	NB	-	-	33.3
ICSI	NB	-	-	27.1
ICSI	WB	-	-	25.3
CTS-ICSI	NB	MAP	1	26.5
CTS-ICSI	NB	MAP	8	25.8
CTS-ICSI	WB	CMLLR + MAP	8	24.6

- But this is not that straight-forward with HLDA, discriminative training, and speaker adaptive training.

Adaptation of CTS Models (2)

- Solution

1. Transform meeting data into NB space
2. Transform full covariance statistics for HLDA and combine with meeting statistics (MAP adaptation)
3. Retrain models in joint HLDA NB space
4. MPE-MAP adapt CTS models to the meeting domain
   Standard MAP adaptation followed by MPE-MAP.

... and include CMLLR based SAT in the process ...

Transformation Between Spaces

- HLDA - based on MAP adapted CTS full-covariance statistics

System	WER [%]
non-adapted WB HLDA system	28.7
HLDA taken from CTS	29.2
HLDA based on adapted statistics	28.1

Results on the RT'05 evaluation set

- Including SAT and discriminative training

1. MPE training of CTS models
2. First adapt using ML-MAP
3. Use models from step 2 as priors for MPE-MAP

Initial models	Adaptation	WER [%]
CTS-SAT-MPE	-	30.4
CTS-SAT-MPE	ML-MAP	26.0
ML-MAP	MPE-MAP	23.9

Results on the RT'05 evaluation set

Putting it together

- Results on CTS with WB/NB adapted HLDA (PLP only, VTLN)

	CTS (270h)	CTS+Fisher (1000h)	CTS+Fisher (2050h)
ML	31.3	29.6	-
MPE	28.0	26.4	25.9

Results on the NIST 2001 Evaluation Set

  - Models trained on 270hours have $\approx 30\%$ fewer parameters

Adapting to meeting data

PLP only	270h	(2050h)
CTS-MPE	30.4	30.4
MAP	26.0	23.8
MPE-MAP	23.9	22.1

Results on the NIST RT'05s Evaluation Set

Meeting data only

	PLP	LCRC+PLP	BN+MFCC
ML	25.8	23.6	23.5
MPE	23.4	21.5	21.5

Results on the NIST RT'05s Evaluation Set

Language Modelling

- N-gram based language modelling

1. UoS Web-data
   - for conference room: 138MW + 54MW
   - lecture room 114MW + 62MW downloaded
2. AMI corpus for RT evals (which excludes the RTxx dev and eval data)
3. CHIL rt06s LM training data
4. CHIL (all Pre- rt07 dev and eval sets merged for LM training)
5. Enron Email
6. Fisher corpus
7. Hub4 Broadcast News 1997
8. ICSI meetings corpus
9. ISL meetings corpus
10. NIST1 and NIST2 meetings corpora
11. Switchboard/Callhome
12. Webdata from UW: Switchboard, Fisher, Fisher topics, Meetings
13. Newly collected webdata for rt07: conf and lect

Web-data Collection

- ▶ Web-data is important for meetings transcription
  1. Large variety of topics
  2. Very limited in-domain data

- ▶ Standard approach (Bulyko, 2003)
  - ▷ Search for N most frequent N-grams in the in-domain data.

- ▶ Search model framework (Wan&Hain, 2006)
  1. Small in-domain corpus T, large background corpus B.
  2. Assume interpolation with existing background corpus
  3. Prediction model for search result
  4. Compute the probability that an n-gram should form a query
     $\Rightarrow$ML optimisation
     The probability for the n-gram (w, h) increases non-linearly with the ratio

$$\frac{P(w|h,T)}{P(w|h,B)}$$

Webdata Collection - Example Queries

Simplest approach: do not use N-grams that already exist in the background material.

Search model approach (trigram)	Using T count only (trigram)
UH REMOTE CONTROL	L. C. D.
TWENTY FIVE EUROS	WE HAVE TO
REMOTE CONTROL UM	I DON'T KNOW
THE SCROLL WHEEL	THE REMOTE CONTROL
OUR REMOTE CONTROL	YOU HAVE TO
NEW REMOTE CONTROL	I THINK WE
USER INTERFACE DESIGNER	I THINK IT'S
THE WORKING DESIGN	THE L. C.
UH THE REMOTE	I DON'T THINK
THE POWER BUTTON	A LOT OF
CHIP ON PRINT	THE T. V.
FASHION IN ELECTRONICS	A REMOTE CONTROL
THE CONCEPTUAL DESIGN	YEAH I THINK
THE MARKETING EXPERT	WE NEED TO
THE FUNCTIONAL DESIGN	YOU WANT TO
FANCY LOOK AND	C. D. SCREEN

Search Model Results on the AMI Corpus

- In-domain data T, evaluation set E, collected corpus C.

Query	C	**Corpus interpolation weights**								PPL	%WER
orders	size	fisher	hub4	swb	icsi	isl	nist	(T)	(C)	on E	on E
—	0	0.21	0.19	0.12	0.27	0.07	0.11	—	—	130.5	35.7
4g	62M	0.18	0.02	0.12	0.21	0.06	0.05	—	0.33	109.3	33.6
3g	61M	0.17	0.01	0.11	0.21	0.05	0.05	—	0.36	100.9	33.1
2g	51M	0.18	0.03	0.11	0.22	0.05	0.05	—	0.32	102.7	33.0
4g+3g	123M	0.17	0.01	0.11	0.21	0.05	0.05	—	0.37	100.8	32.9
4g+3g+2g	174M	0.16	0.01	0.11	0.21	0.05	0.05	—	0.38	94.2	32.2
—	0	0.10	0.09	0.02	0.03	0.00	0.00	0.72	—	76.9	32.2
4g	62M	0.09	0.02	0.01	0.02	0.00	0.00	0.69	0.14	74.4	31.9
3g	61M	0.09	0.02	0.01	0.02	0.00	0.00	0.67	0.16	72.6	31.8
2g	51M	0.05	0.04	0.00	0.00	0.00	0.00	0.68	0.20	73.3	31.9
4g+3g	123M	0.08	0.01	0.01	0.01	0.00	0.00	0.67	0.17	72.2	31.7
4g+3g+2g	174M	0.08	0.01	0.01	0.01	0.00	0.00	0.67	0.18	69.1	31.2

System Overview

- IHM: Four acoustic model sets

  M1 models: PLP features, MPE

  M2 models: PLP + LCRC features, VTLN, SAT, MPE, meeting data only

  M3 models: MFCC + Bottleneck, VTLN, SAT, MPE, meeting data only

  M4 models: PLP, VTLN, HLDA, SAT, NB/WB, MPE-MAP adapted from CTS/Fisher models

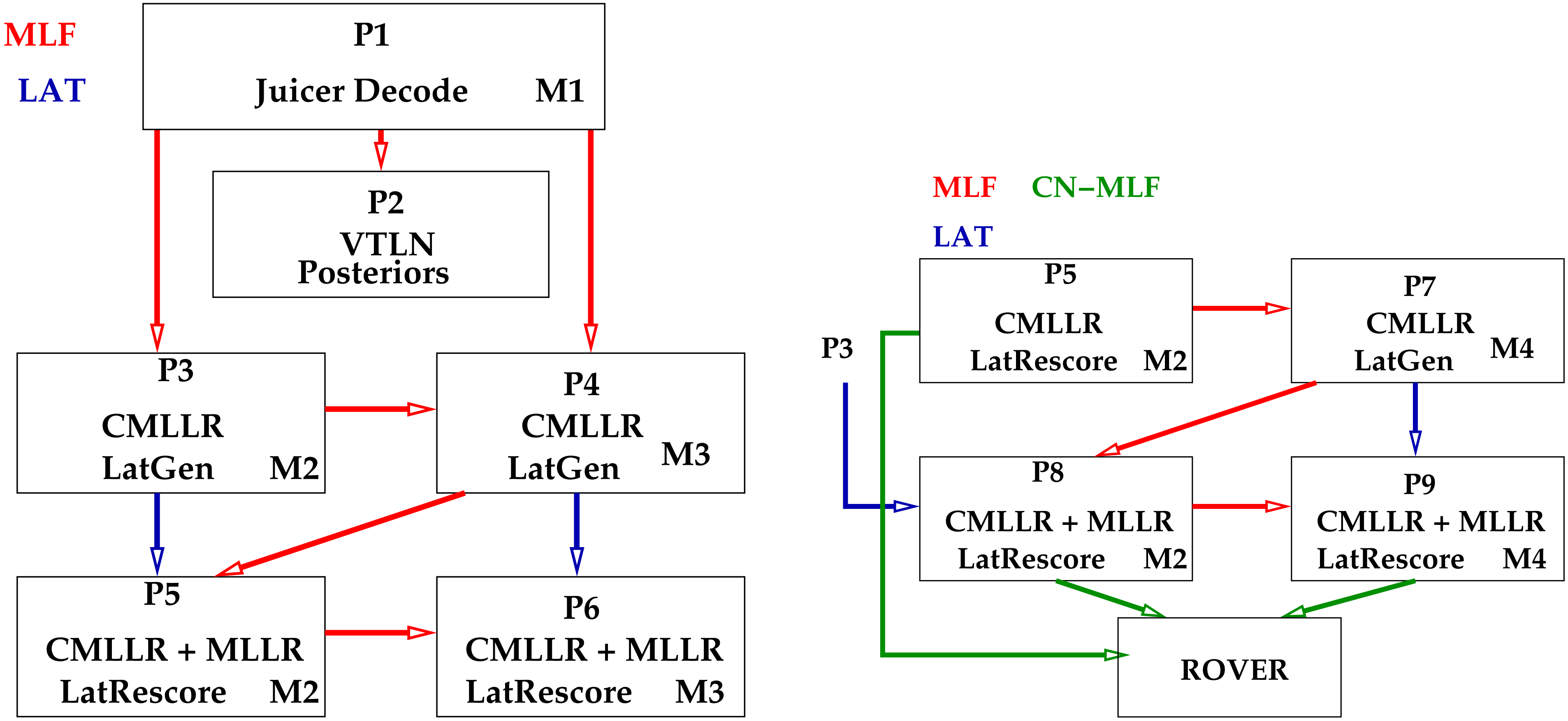
- MDM:

  Only M1 and M2 models were trained

- Bigram, Trigram, 4-gram language models

- Additional features: Confusion networks, ROVER

2007 System Architecture



2007 Performance Conference Meeting

- Individual Head Microphone
- *RT'06 Evaluation Set* (10 minute extracts from 8 meetings).

-	TOT	Sub	Del	Ins	CMU	EDI	NIST	TNO	VT
P1	35.4	19.3	12.8	3.2	35.4	32.5	31.5	35.2	39.8
P3	24.9	12.8	9.7	2.5	24.9	23.0	22.4	25.0	29.3
P4	24.4	12.4	9.6	2.4	24.4	22.7	21.7	23.9	28.8
P5	23.7	11.8	9.7	2.2	23.7	21.9	21.1	24.2	27.9
P5.cn	23.4	11.7	9.6	2.1	23.4	21.6	20.8	24.0	27.8
P6	23.7	11.9	9.5	2.3	23.7	21.6	21.3	24.0	28.0
P6.cn	**23.5**	11.7	9.5	2.3	23.5	21.7	21.0	23.9	27.7
P7	24.1	12.5	9.2	2.4	24.0	22.8	22.2	22.4	28.7
P8	23.2	11.7	9.2	2.2	23.2	21.3	20.9	22.8	27.7
P8.cn	22.9	11.6	9.1	2.2	22.9	21.1	20.7	22.5	27.3
P9.cn	23.7	12.2	9.2	2.4	23.6	22.4	21.9	22.2	27.9
final.rover	**22.3**	11.0	9.3	2.0	**22.2**	**20.7**	**20.2**	**22.1**	**26.7**

2007 Performance Conference Meeting

- Individual Head Microphone
- *RT'07 Evaluation Set*

	TOT	Sub	Del	Ins	CMU	EDI	VT
P1	37.4	20.6	12.9	4.0	41.5	28.4	41.3
P3.fg	28.2	14.5	10.4	3.3	33.7	19.8	30.8
P4	27.9	14.1	10.6	3.2	33.1	20.0	30.2
P5	27.7	13.5	11.1	3.1	34.5	19.5	30.4
P5.cn	25.9	13.5	9.9	2.5	31.2	18.3	28.5
P6.cn	25.7	13.6	9.5	2.6	30.6	18.4	28.2
P7	27.9	14.5	9.9	3.4	34.7	20.3	29.6
P8	26.9	13.6	10.1	3.3	32.0	19.4	29.6
P8.cn	25.4	13.4	9.4	2.6	30.8	18.0	27.2
P9	27.9	14.6	9.9	3.5	34.7	20.4	29.6
P9.cn	26.3	14.3	9.3	2.7	33.5	19.0	27.1
P5+P8+P9	**24.9**	**12.7**	**9.8**	**2.4**	**30.5**	**17.6**	**26.8**

2007 Performance Conference Meeting - Manual segmentation

- Individual Head Microphone
- *RT'07 Evaluation Set*

	TOT	Sub	Del	Ins	CMU	EDI	VT
P1	34.2	21.7	10.0	2.6	38.3	25.3	38.9
P3.fg	25.2	15.5	7.6	2.1	30.6	16.8	28.6
P4	24.5	15.0	7.5	2.0	29.0	16.8	27.4
P5	24.1	14.9	7.4	1.9	28.8	16.3	27.7
P5.cn	23.8	14.6	7.3	1.8	28.0	16.1	27.4
P6.cn	**23.6**	**14.5**	**7.1**	**2.0**	**27.4**	**16.3**	**27.4**
IHM P6.cn	**25.7**	**13.6**	**9.5**	**2.6**	**30.6**	**18.4**	**28.2**

2007 Performance Conference Meeting

- Multiple Distant Microphones
- *RT'07 Evaluation Set*

	ICSI S&C				AMI/DA S&C			
	TOT	Sub	Del	Ins	TOT	Sub	Del	Ins
P1	44.2	25.6	14.9	3.8	44.7	25.7	16.3	2.7
P3	38.9	18.5	16.8	3.5	34.5	19.3	12.5	2.7
FINAL	33.7	20.1	10.7	2.9	33.8	19.2	12.2	2.4
FINAL manual seg	30.2	18.7	9.4	2.0	-	-	-	-

We thank ICSI for providing segmentation and clustering (S&C).

Conclusions

- Presented the AMI 2007 Meeting System
  - Competitive performance
  - Considerable gap between MDM and IHM.
- Good results for
  - IHM Segmentation
  - Feature extraction
  - Webdata
- Improvement needed
  - MDM front-end
  - Speed
- Meeting transcription is a still difficult task
  - Both IHM and MDM front-ends need improvement
  - Participation in NIST RT evaluations is low ...

Selected References

- S. Fitt (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules, Tech. Rep., Centre for Speech Technology Research, Edinburgh.
- J. Dines, J. Vepa and T. Hain (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In. Proc. Interspeech 2006.
- T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa and V. Wan (2007). The AMI system for the Transcription of Speech in Meetings, In Proc ICASSP 2007. 15-20 April 2007, Honolulu, Hawaii, USA.
- M. Karafiat, L. Burget, J. Cernocky, T. Hain (2007). Application of CMLLR in NB-WB adapted systems. To appear in Proc. Interspeech 2007.
- P. Schwarz, P. Matjka and J. Cernock (2004), Towards Lower Error Rates in Phoneme Recognition, in Proc. of 7th Intl. Conf. on Text, Speech and Dialogue, no. ISBN 3-540-23049-1, pp. 8, 2004
- V. Wan and T. Hain (2006). Strategies for Language Model Web-data Collection. Wan, V. & Hain, T., Proc. ICASSP'06