AMI/DA STT and SASTT 2007

Thomas Hain - UoS

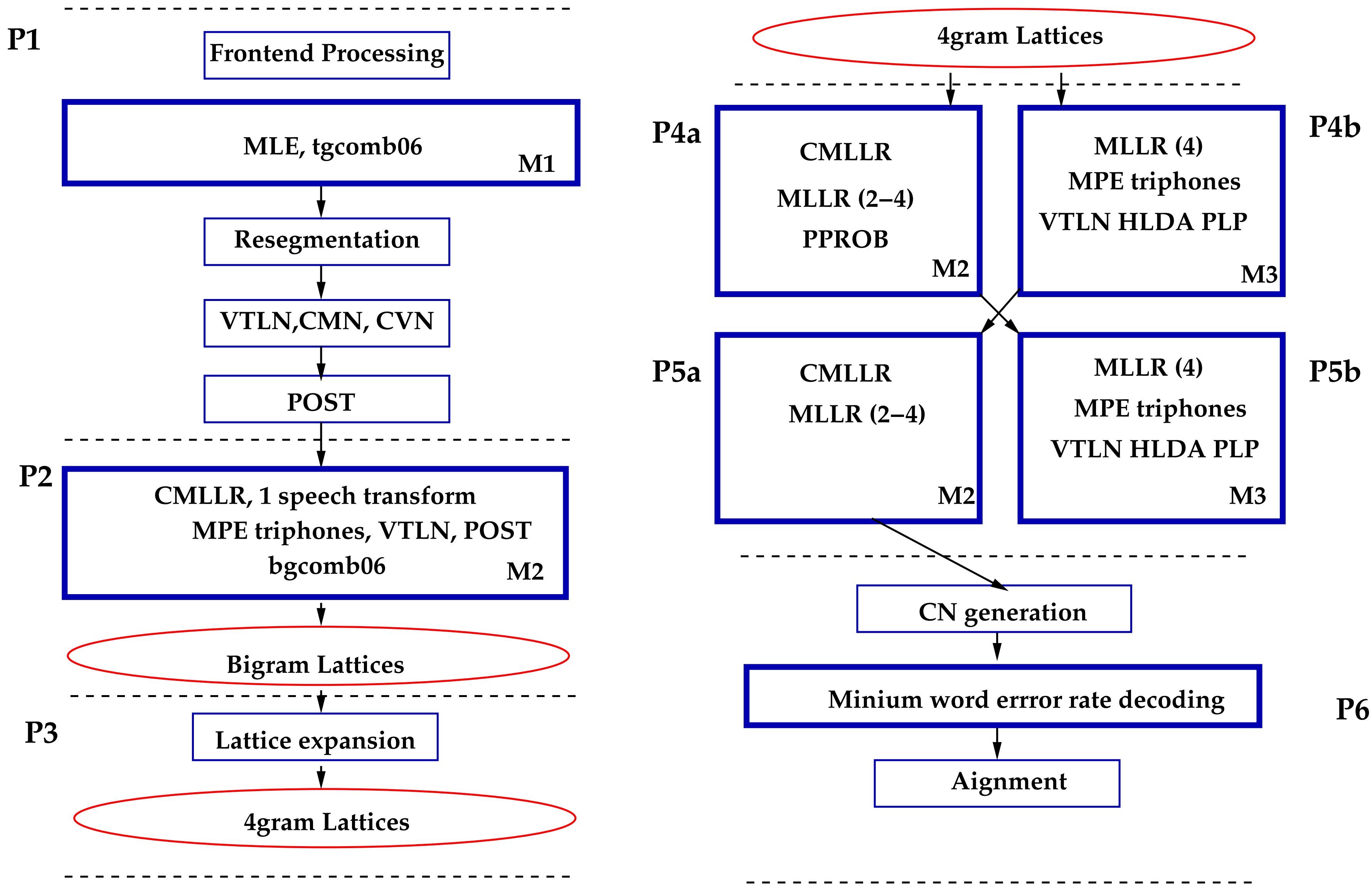
Lukas Burget, Martin Karafiat - BUT
John Dines - IDIAP
David van Leeuwen - TNO
Giulia Garau, Mike Lincoln - Univ Edinburgh
Vincent Wan- UoS

May 10, 2007

Outline

1. Review of the 2006 System
2. New in the 2007 System
3. Features not yet included
4. Summary

2006 - System architecture



2006 - Main features

- Acoustic modelling
  - 3 models
  - Posterior features : LCRC
  - WB/NB adaptation including SAT and MPE
- Language modelling
  - Search model based LM data collection and use
- Decoding
  - New FST decoder for first pass
- Improved IHM front-end

Results RT06S - Conference

IHM	TOT	Sub	Del	Ins	CMU	EDI	NIST	TNO	VT
P1	**42.0**	25.3	12.6	4.1	41.9	41.0	39.0	42.1	44.8
P3	**26.0**	13.9	9.5	2.6	25.7	24.6	25.2	26.3	29.5
P4a	**25.1**	13.0	10.0	2.1	25.0	22.8	23.8	26.0	29.1
P4b	**25.6**	13.3	10.2	2.1	25.3	23.8	24.9	24.3	29.8
P5a	**24.6**	12.6	10.0	2.0	24.4	22.6	23.6	24.1	28.8
P5b	**27.6**	12.8	12.8	2.0	27.1	26.7	31.3	24.2	29.8
P5a-cn	**24.2**	12.3	10.0	1.9	24.0	22.2	23.2	23.6	28.2

MDM	TOT	Sub	Del	Ins
P1	**58.2**	35.8	16.7	5.7
P2a	**45.6**	26.4	15.1	4.1
P3	**42.0**	24.5	13.2	4.4
P4a	**41.7**	22.9	14.9	3.9
P4a-CN	**40.9**	22.2	15.3	3.5

2006 - Lecture

General strategy (since 2005)

- Use the conference acoustic models and system architecture
- Use domain specific LMs

IHM	Segmentation	TOT	Sub	Del	Ins
P1	auto	81.8	31.7	7.4	42.7
P5a-CN	auto	57.8	18.2	7.3	32.2
P1	manual	50.4	31.7	7.0	11.7

MDM	TOT	Sub	Del	Ins
P1	71.4	47.5	14.4	9.5
P4a-cn	58.1	28.7	23.9	5.5

New in the 2007 System

1. Dictionary and word list expansion and cleaning
2. New training data (and hence new models)
   (a) *ihmtrain07* and *mdmtrain07*: includes new NIST and AMI data
   (b) *ctstrain07*: now includes 2000 hours of Fisher data
3. LM optimisation routine
4. IHM segmentation optimisation
5. Included AMI MDM segmentation and clustering
6. Alternative front-end: MFCC + Bottleneck features
7. SASTT
8. ROVER / CNC
9. System architecture
10. Coffee break

Not quite made it

Either discouraging results or not ready:

1. MLP based LMs
2. Window based MLLR
3. STRAIGHT features
4. CHAT
5. New system development software

Dictionary and Wordlist

- Baseline dictionary based upon UNISYN (Fitt, 2000) with ∼115,000 words
- Prior to 2006 NIST evaluations we had added ∼11,500 words
- This year, additional words added:
  - ∼1,750 to give full coverage of AMI corpus (including part words)
  - ∼1,500 for the Fisher corpus

BUT

- Word list problems (increased conceivability)
  - with compound words
  - hyphenated words
  - acronyms
  - partial words

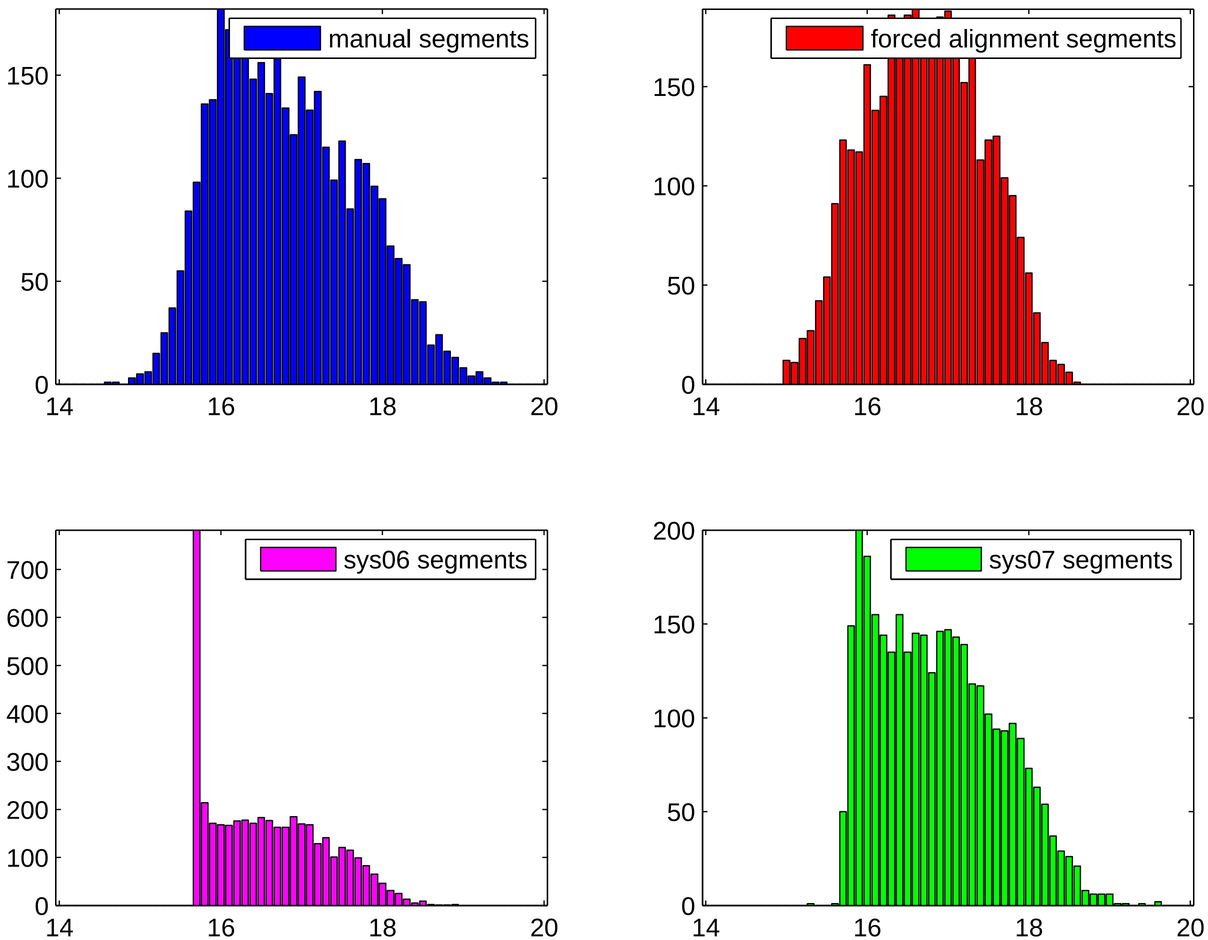
Word list quality checklist

- All words are classified according to a quality ranging from 1- 5
  - Lowest quality words may contain illegal symbols
  - Highest quality are words either correct in spell check or manually checked
- Words are assigned initial quality
  - part-words are Q1
  - single letters are Q5
  - words with highest spell-check level are Q5
  - ...
- Rule based selection of candidates for manual check
- Only words higher than Q3 may be included in test dictionaries

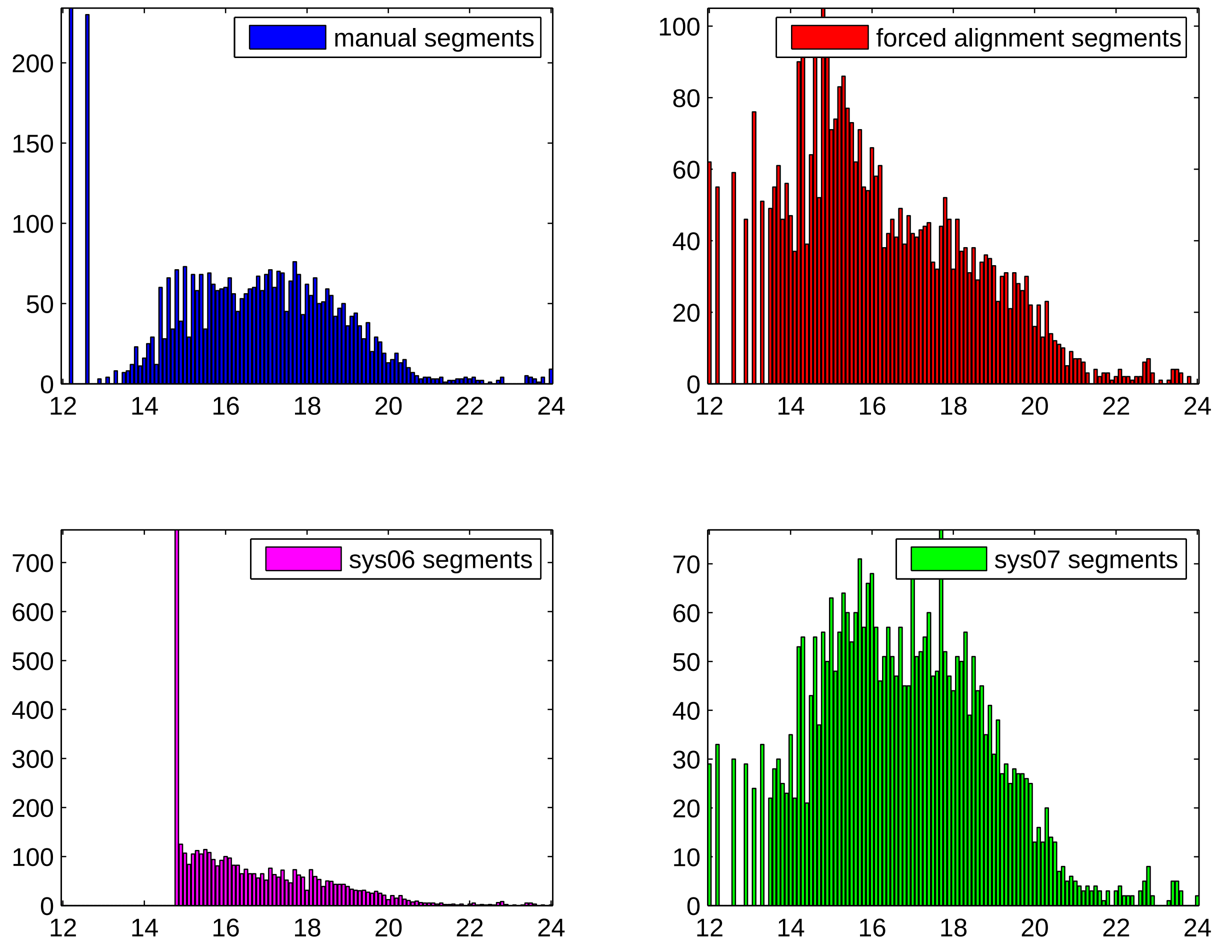
IHM Front-end Speech Activity Detection

- Same approach as used as in 2006 for the classifier:
  - Multilayer perceptron (MLP) consisting of 15 input frames (of feature dimension 51), 50 hidden units and 2 output classes (speech/non-speech) trained from forced alignment of 150 meetings
  - Speech activity detection using a Viterbi decoding of scaled likelihoods from MLP + 200 ms silence collar
- Tuning of hyper-parameters (min state duration, segment insertion penalty, silence collar) carried out by matching with duration histograms for ground truth segmentations
  - In RT06, hyper-parameters were chosen to be *sensible*' values
  - In RT07 hyper-parameters tuned based on segment duration histograms from manual segmentations

Histogram of speech segment durations



Histogram of silence segment durations log(100 ns)



MDM Audio processing

- Essentially the same as RT06
- For all rooms with ¡2 omni directional microphones
  - Wiener filter static noise removal
  - GCC delay estimate
  - Frequency domain super-directive beamformer
- For NIST room (4 way directional mic)
  - Noise removal
  - Select highest energy channel on a per frame basis (0.5 seconds)
  - Used data from 4 way directional mic

CHIL recordings:

- Convert all files to 16bit 16Khz (UPC files in particular)

Optimisation of speaker segmentation/clustering

- Speaker segment/clusterer has several purposes:
  - Speech activity detection, limiting segment duration for decoder
  - Speaker adaptation: VTLN, CMLLR
  - Diarisation (SASTT)
- Parameter settings not the same for these purposes

	#clusters	WER (%)	DER (%)
ICSI reference	4	56.8	-
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

rt06seval

- Better results for 1st pass WER were obtained with simple BIC segmentation/clustering (AMI 2006 system) than current SPKR system.

Language modelling - Data

1. UoS Webdata
   (a) for conference room: 138MW + 54MW
   (b) 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

AMI corpus - Analysis

Collection	Overall	male	female	Scenario	Non-Scen
cts+bn+meetings+wuweb001	92.929	92.826	93.176	84.118	119.667
wuweb001	94.324	94.180	94.676	84.497	124.728
cts+bn+meetings001	96.367	96.064	97.100	86.537	126.702
bn001	99.801	99.322	100.937	87.914	137.806
cts001	100.507	100.057	101.592	88.204	140.146
meetings001	102.701	101.587	105.390	91.242	138.817

Collection	English	French	German	OtherEU	S. Asia	Rest of World
cts+bn+meetings+wuweb001	96.923	90.803	110.998	103.032	104.718	94.905
wuweb001	98.585	92.156	113.848	106.518	107.372	96.875
cts+bn+meetings001	101.110	94.279	119.108	106.693	107.757	98.618
bn001	105.188	97.676	128.481	113.289	111.981	102.802
cts001	105.903	100.223	128.890	114.384	114.968	103.963
meetings001	110.332	97.959	126.765	115.941	113.286	103.702

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

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

LP LMs trained on meeting corpora and Fisher corpus

Combining MLP LMs (Schwenk 2004) with latent semantic analysis (LSA)

- Top 6800 words
- LSA on Gigaword corpus to yield 200D vector
- 4gram MLP
  - LSA vectors represent words (thus 600 inputs in total);
  - 300 hidden units gave the best reduction in perplexity;
  - 6800 output layer
- OOV words produce zero vector

hidden units	conf PPL	lect PPL
50	81.79	143.41
100	78.25	138.57
150	78.20	140.16
200	77.60	138.71
250	77.38	138.78
300	76.93	137.74

Perplexities on the rt07s LM development set

LM training procedure

- STAGE1
  - Take 9 most highly weighted language models
  - Use in search model framework to rank 4grams in the texts of ”the RT evals previous to RT07sdev”.
  - Use top 600-2000 queries from several search models to collect 20MW
  - train additional LM component
- STAGE2
  - LMs reconstructed by interpolating the 10 most highly weighted LMs from the total list
  - No component LM with interpolation weight < 0.01
  - new web-data dropped out !
- STAGE3
  - Interpolate with MLP

Conference LMs - interpolation weights

corpus	weight
fisher webdata from UW	0.220
amicorpus4RTevals	0.210
fisher-03	0.186
meetings webdata from UW	0.103
isl-mc1	0.081
switchboard+callhome	0.048
swb webdata from UW	0.045
amicorpus webdata	0.038
hub4lm96	0.035
nist-2	0.029

STAGE1

Perplexity on *rt06seval*: 73.2

corpus	weight
P4 conf LM	0.912
rt06s conf webdata	0.054
icsi	0.019
nist-all	0.014

STAGE2

Perplexity on *rt06seval*: 73.1

- New web-data collection dropped out: We have all the data available .... (?)

Including MLPs

For lecture room data

- Largest weights on STAGE1 model: meeting WEBDATA , CHIL, ICSI, AMI , ...
- STAGE2: 11% on rt06s lectmtg webdata collection

Including MLPs

rt06seval	STAGE2	MLP	STAGE2 weight	MLP weight	Combined
confmtg	73.1	76.9	0.90	0.10	72.7
lectmtg	119.3	137.7	0.92	0.08	118.1

Cross domain LM testing

4-gram LM	*confmtg* (*rt06seval*)	*lectmtg* (*rt07slmdev*)
RT06 LM	75.2	125.8
confmtg STAGE1	73.2	144.5
confmtg STAGE2	73.1	140.8
lectmtg STAGE1	82.9	120.4
lectmtg STAGE2	81.9	119.3

- Optimisation for coffee break data was deemed unnecessary with perplexities around 95.

rt07seval	TOT	Sub	Del	Ins	CMU	EDI	NIST	VT
lm05	28.7	14.9	10.2	3.5	33.6	20.7	14.7	31.7
lm06	28.6	14.9	10.2	3.5	34.1	20.2	14.4	31.5
lm07	28.5	14.8	10.2	3.5	34.0	20.3	14.4	31.1

Language model (trigram) and dictionary change

Acoustic training data

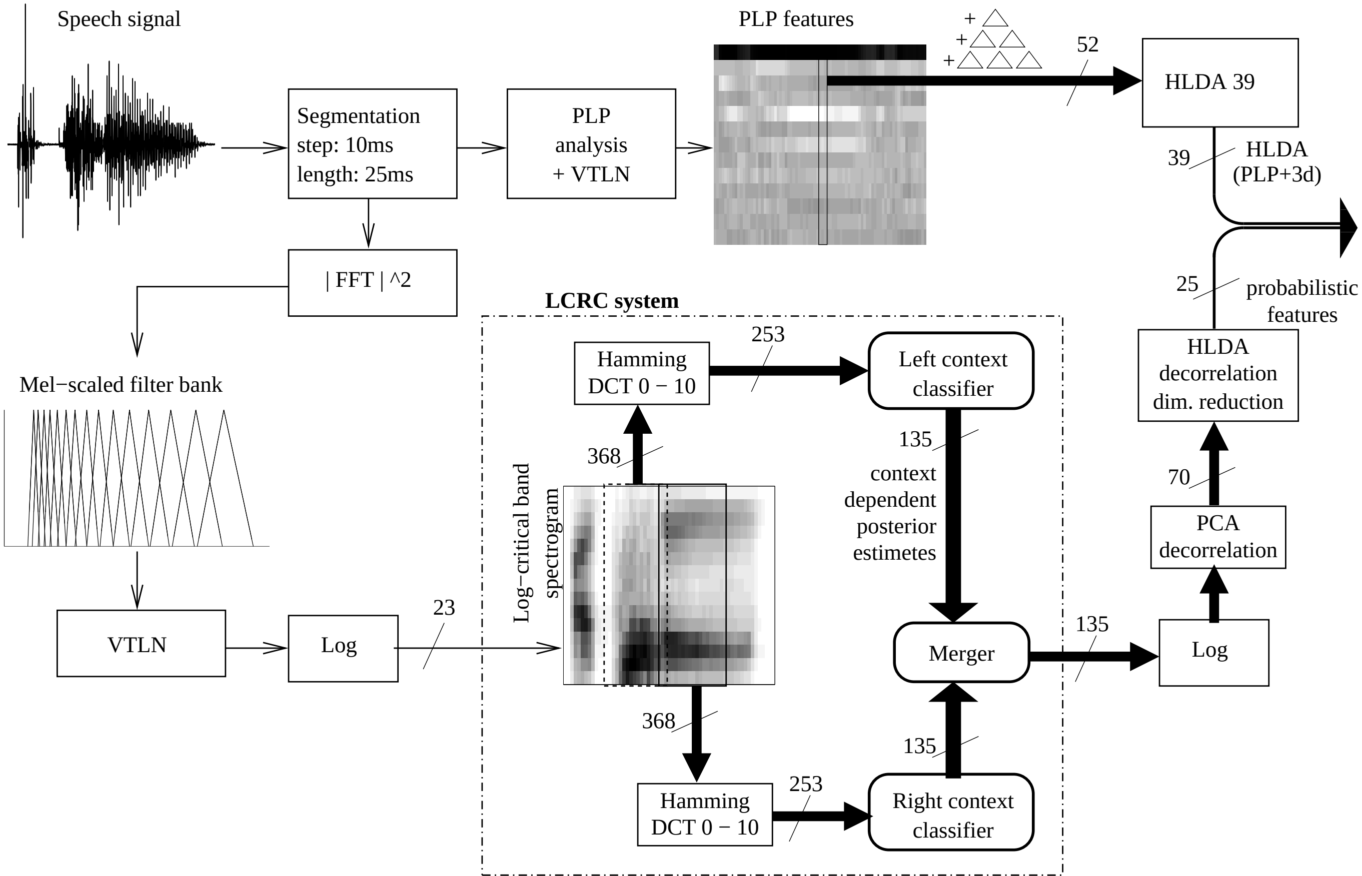
- IHM training data: *ihmtrain07* includes new NIST and AMI data

  172.89 hours, excluded 8 hours of silence

- MDM training data: *mdmtrain07* selected to exclude overlap (x in WB x denotes minimum distance from word boundary).

	#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

Posterior features



- ▶ MLPs are now trained on 100 hours of speech
- ▶ Bottleneck features 25 dimensional on single MLP for complete spectrum

Meeting models

M2 models: PLP + LCRC features, trained on meeting data only

M3 models: MFCC + Bottleneck, trained on meeting data only

	Features	Tr	Adapt/Normalise	TOT	CMU	EDI	NIST	TNO	VT
M2	PLP	ML		39.0	39.0	35.4	33.7	40.3	45.6
	PLP	ML	VTLN HLDA	31.8	31.9	29.0	29.1	30.0	37.9
	PLP + LCRC	ML	VTLN HLDA	-	-	-	-	-	-
	PLP + LCRC	ML	VTLN HLDA SAT	27.2	27.2	25.0	25.0	27.1	32.1
	PLP + LCRC	MPE	VTLN HLDA SAT	25.4	25.4	23.3	23.3	25.2	29.4

	Features	Tr	Adapt/Normalise	TOT	CMU	EDI	NIST	TNO	VT
M3	MFCC	ML		39.7	39.9	37.0	34.2	38.9	45.8
	MFCC	ML	VTLN HLDA	34.2	34.2	32.6	29.9	32.0	41.0
	MFCC + BN	ML	VTLN HLDA	29.4	29.3	27.5	26.6	28.1	35.6
	MFCC + BN	ML	VTLN HLDA SAT	27.3	27.2	25.2	25.6	26.5	32.3
	MFCC + BN	MPE	VTLN HLDA SAT	25.6	25.6	23.0	23.6	24.9	30.1

Meeting models : MDM

	Features	Tr	Adapt/Normalise	TOT	Sub	Del	Ins
M2	PLP	ML		53.7	33.6	14.0	6.1
	PLP	ML	VTLN HLDA	48.0	27.7	14.0	6.2
	PLP + LCRC	ML	VTLN HLDA	42.8	25.6	11.2	6.0
	PLP + LCRC	ML	VTLN HLDA SAT	40.9	24.1	12.4	4.4
	PLP + LCRC	MPE	VTLN HLDA SAT	37.9	21.9	12.0	4.0

Adaptation to the meeting domain

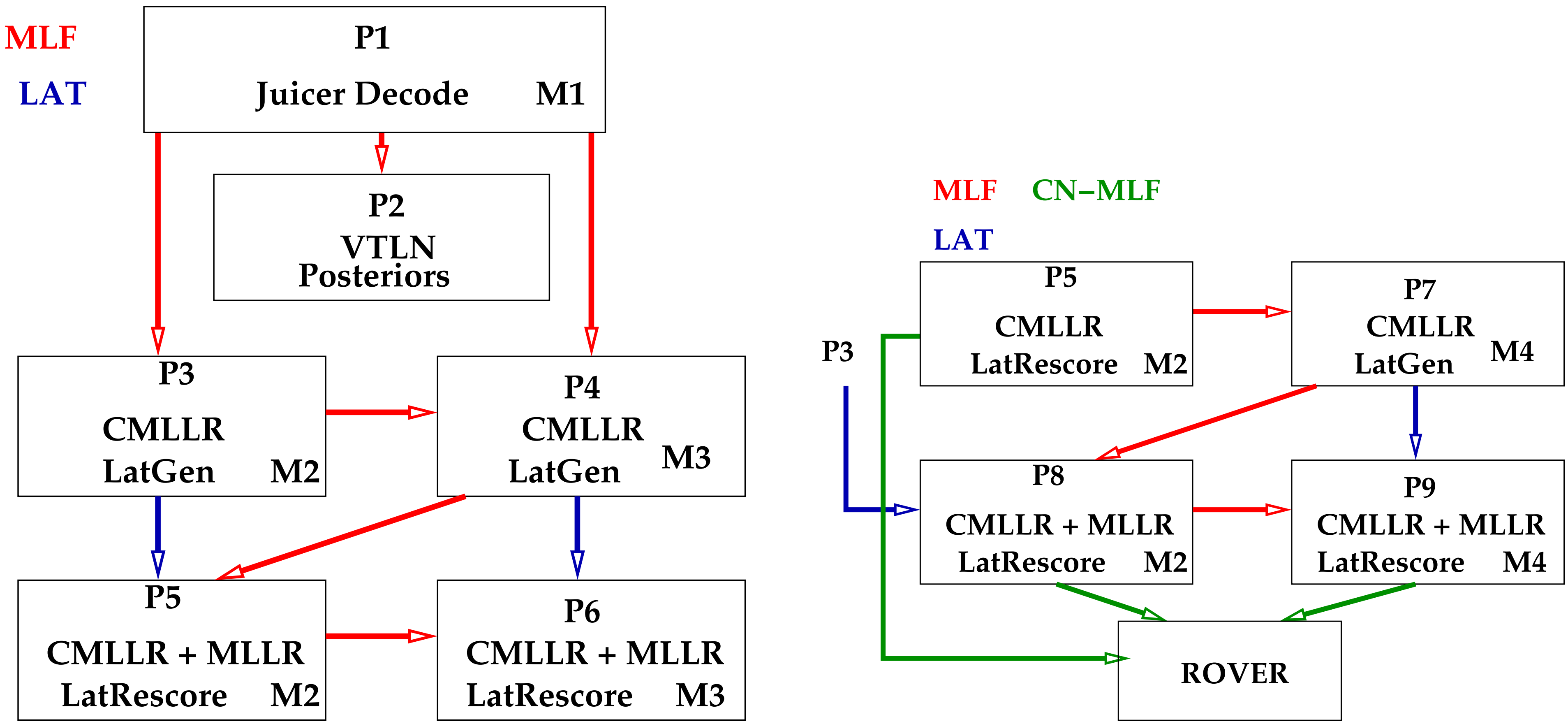
- Motivation
  - Smoothing due to substantial increase of training data
- Issues:
  - Narrowband (NB) vs Wideband (WB)
  - HLDA statistics collected on more data
- Solution
  1. Transform meeting data into NB space
  2. Transform full covariance statistics for HLDA and combine with meeting statistics (MAP adaptation)
  3. Retrain models in joint HLDA NB space
  4. MPE-MAP adapt CTS models to the meeting domain

... and include SAT in the process ... ⇒***M4 models***

Fisher adapted models

- Fisher corpus
  - Fisher + CTS = 2300 hours, removed 175 hours of silence
  - Excluded all segs with words occurring less than 4 times
- Fisher models
  - ML training on 1000 hours
  - 10k states, 20 Gaussians per state
  - $\approx 1\%$ better performance on CTS compared to CTS only trainig
  - MPE training on 2000 hours (no posteriors !)

2007 System Architecture



2007 Performance Conference Meeting -IHM - *rt06seval*

-	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 -IHM - *rt07seval*

	TOT	Sub	Del	Ins	CMU	EDI	NIST	VT
P1	37.4	20.6	12.9	4.0	41.5	28.4	18.8	41.3
P3.fg	28.2	14.5	10.4	3.3	33.7	19.8	14.1	30.8
P4	27.9	14.1	10.6	3.2	33.1	20.0	13.8	30.2
P5	27.7	13.5	11.1	3.1	34.5	19.5	13.6	30.4
P5.cn	25.9	13.5	9.9	2.5	31.2	18.3	12.0	28.5
P6.cn=final	**25.7**	**13.6**	**9.5**	**2.6**	**30.6**	**18.4**	**11.8**	**28.2**
P7	27.9	14.5	9.9	3.4	34.7	20.3	13.9	29.6
P8	26.9	13.6	10.1	3.3	32.0	19.4	13.3	29.6
P8.cn	25.4	13.4	9.4	2.6	30.8	18.0	11.7	27.2
P9	27.9	14.6	9.9	3.5	34.7	20.4	14.0	29.6
P9.cn	26.3	14.3	9.3	2.7	33.5	19.0	12.3	27.1
P5+P8+P9	**24.9**	**12.7**	**9.8**	**2.4**	**30.5**	**17.6**	**11.5**	**26.8**

2007 Performance Conference Meeting -IHMREF - *rt07seval*

- Raw manual segmentation (no alignment)

	TOT	Sub	Del	Ins	CMU	EDI	NIST	VT
P1	34.2	21.7	10.0	2.6	38.3	25.3	16.4	38.9
P3.fg	25.2	15.5	7.6	2.1	30.6	16.8	11.5	28.6
P4	24.5	15.0	7.5	2.0	29.0	16.8	11.0	27.4
P5	24.1	14.9	7.4	1.9	28.8	16.3	11.0	27.7
P5.cn	23.8	14.6	7.3	1.8	28.0	16.1	10.9	27.4
P6.cn	**23.6**	**14.5**	**7.1**	**2.0**	**27.4**	**16.3**	**10.8**	**27.4**
IHM P6.cn	**25.7**	**13.6**	**9.5**	**2.6**	**30.6**	**18.4**	**11.8**	**28.2**

- Also tested automatic res-segmentation of data $\Rightarrow$poorer Performance

2007 Performance Conference Meeting - *rt07seval* - MDM

	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	-	-	-	-

- Substantial differences between segment's
  - Performance level may hide weaknesses
- Manual segmentation substantially better
- 37.1% on *rt06seval*

2007 Performance Lectures

- STT performance

STT	TOT	Sub	Del	Ins
P1	61.4	36.4	16.0	9.1
P3	51.0	29.5	13.4	8.1
FINAL	48.2	30.1	12.0	6.1

- SASTT based on optimisation for STT !
  - SASTT performance 51.5%

Window-based MLLR

- MDM channels are time variant: speakers move around and the acoustic conditions tend to change
- Moving window to estimate the MLLR transforms
  - *transform estimation*: the start of the segment must be simply inside a window
  - *decoding*: the same ...
- Parent CMLLR transform estimated using the whole channel
- In 64 dimensional features space data sparsity seems to be a challenging issue
- Also tried to substitute the “whole channel” transform when no transform could be estimated using the data in the local window

Windows based MLLR: results

rt05seval - MDM	TOT	AMI	CMU	ICSI	NIST	VT
A) global full CMLLR+full MLLR	34.5	30.5	33.2	36.6	37.2	35.3
B) global full CMLLR+2 min width 1min shift full MLLR	34.8	31.2	32.6	37.2	37.2	35.7
C) global full CMLLR+4 min width 1min shift full MLLR	34.8	30.9	32.9	37.3	37.3	35.7
D) global full CMLLR+2 min width 1min shift diag MLLR	35.3	31.9	35.0	36.8	36.9	36.2
E) like B) subs. glob. when ¡ 1xform	34.7	31.1	32.3	36.8	37.3	35.6
F) like B) subs. glob. when ¡ 2xform	34.6	30.9	32.3	36.8	37.3	35.6

Conclusions/Summary

- improvement on both IHM and MDM
  - IHM front-end performance reasonable but could be improved
  - Successful integration of MDM front-end
  - Extended system architecture not fully exploited
  - Several promising approaches in develpoment
- Scoring of SASTT ....
- **THANKS**
  - All people in AMI/DA for helping with getting our system together
  - ICSI/SRI for their continued support providing MDM segmentation and speaker information