The AMI Meeting STT System - Release 2

Thomas Hain - UoS.

Lukas Burget, Martin Karafiat - BUT
John Dines, Jithendra Vepa - IDIAP
Giulia Garau,Mike Lincoln - Univ Edinburgh
Vincent Wan- UoS.

May 3, 2006

Outline

1. Review of the 2005 System

2. What is new in 2006

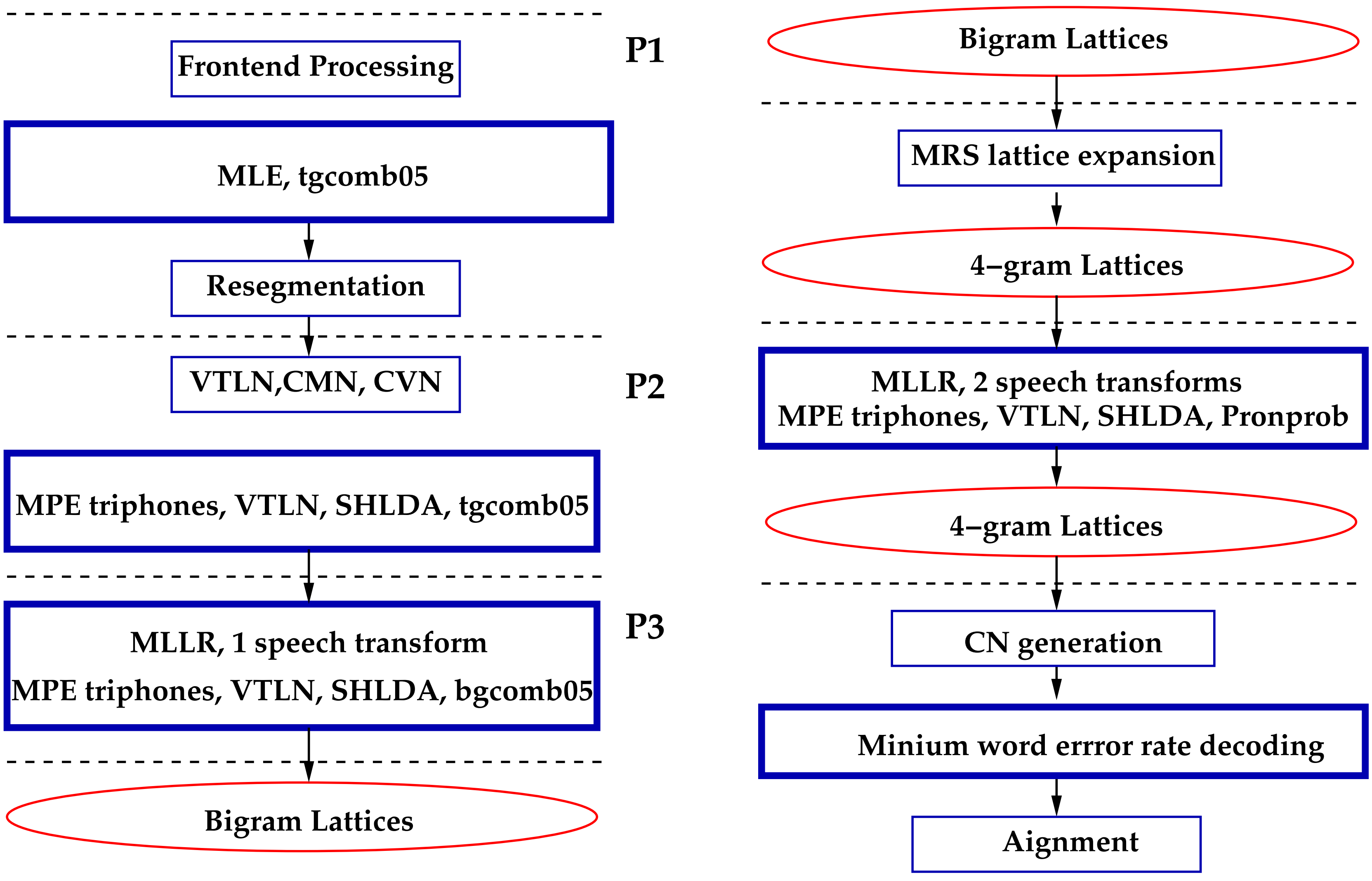
   (a) Front-ends
   (b) Language modelling
   (c) Posterior features
   (d) Acoustic modelling

3. Things that did not make it

4. System architecture and results

5. Summary

Review of the AMI 2005 System



Results and Issues

Key features

- Unisyn dictionary
- SHLDA
- discriminative training
- web-data collection for LMs
- speaker adaptive

Results RT05

	TOT	Sub	Del	Ins	Fem	Male	AMI	ISL	ICSI	NIST	VT
IHM	30.6	14.7	12.5	3.4	30.6	25.9	30.9	24.6	30.7	37.9	28.9
MDM	42.0	25.5	13.0	3.5	42.0	42.0	35.1	37.1	38.4	41.5	51.1

Also tested on lecture room data ...

The 2006 System Development

- New forms of computer failure !

- Things that made it
  - Improved front-ends
  - Improved HLDA
  - Posterior features
  - SAT
  - CMLLR/MLLR adaptation
  - Acoustic feature space mappings / MAP adapted HLDA
  - Search model based LM data collection
  - Faster initial pass - Juicer
  - Modified system architecture

- Things that did not make it
  - Dictionary mappings
  - CML-LR
  - Windowed adaptation
  - CHAT
  - LM adaptation
  - CNC

IHM Front-end

- Adaptive LMS based signal cross-talk suppression
- **Features**: 13 MF-PLP + energy / Cross-channel normalised energy / Signal kurtosis
- **MLP classifier:** 31 input frames 2 output classes
- **Segmentation** Segment minimum duration of 0.5 seconds, enforced via Viterbi decoding of scaled likelihoods Added 0.1 second silence collar to segments

Changes

2005

- Training
  - 20 hrs / 10 hours validation
  - equally sampled from 4 corpora
- Features
  - ZCR
  - Cepstrum based voicing strength
  - 36D (inc differentials)
- MLP
  - 5 hidden units (7k parameters)
  - Priors obtained from training data

2006

- Training
  - 90 hours / 10 validation
  - from all meetings
- Features
  - Maximum normalised cross-correlation
  - Mean cross-correlation
  - 54D (1st and 2nd order differentials)
- MLP
  - 50 hidden units (58k parameters)
  - Priors obtained from RT05s

IHM Front-end - RT06 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

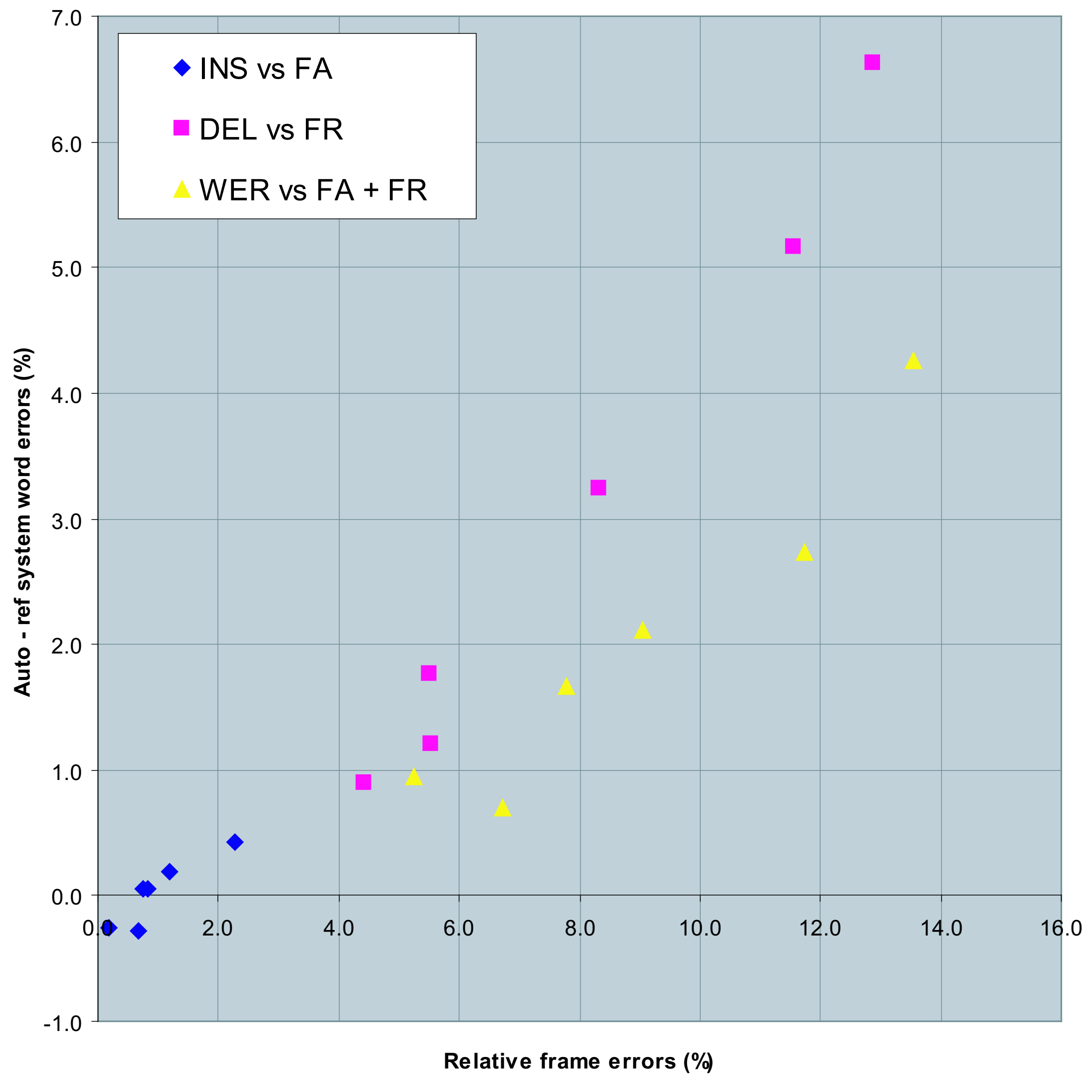
	TOT
INS	**3.5**
DEL	**9.4**
SUB	**26.5**
WER	**39.3**

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

Relationship - Frame error / WER



MDM Processing - 2005

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. **Beamforming:** Frame based frequency domain filtering

$$\mathbf{d}(f) = \left[\alpha_1 e^{-2\pi f \tau_1}; \alpha_2 e^{-2\pi f \tau_2}, \ldots\right]$$

Segmentation and Speaker Clusters again provided by ICSI/SRI.

Changes

- System performs badly on Virginia Tech. recordings
  - Only 2 microphones, widely spaced
- Solution: In cases with 2 microphones, simply pick highest energy channel for every time frame
- And some bug fixes ...

System	Total	AMI	CMU	ICSI	NIST	VT
2005	49.1	41.3	48.0	43.4	50.3	57.9
2006	46.9	41.5	46.6	43.4	49.1	51.8

LM: New Web-data Collection

- RT05s web-data collection:
  - Collected using 4g queries that did not occur in existing corpora
    - 78MW for conference room meetings
    - 68MW for lectures
- New RT06s web-data collection
  - Collected using 3g and 4g queries using the search model framework
    - 60MW for conference room meetings
    - 46MW for lectures
  - RT06s collections were combined with the RT05s collections
    - 138MW in total for conference room meetings
    - 114MW total for lectures
- Minor improvements in perplexity

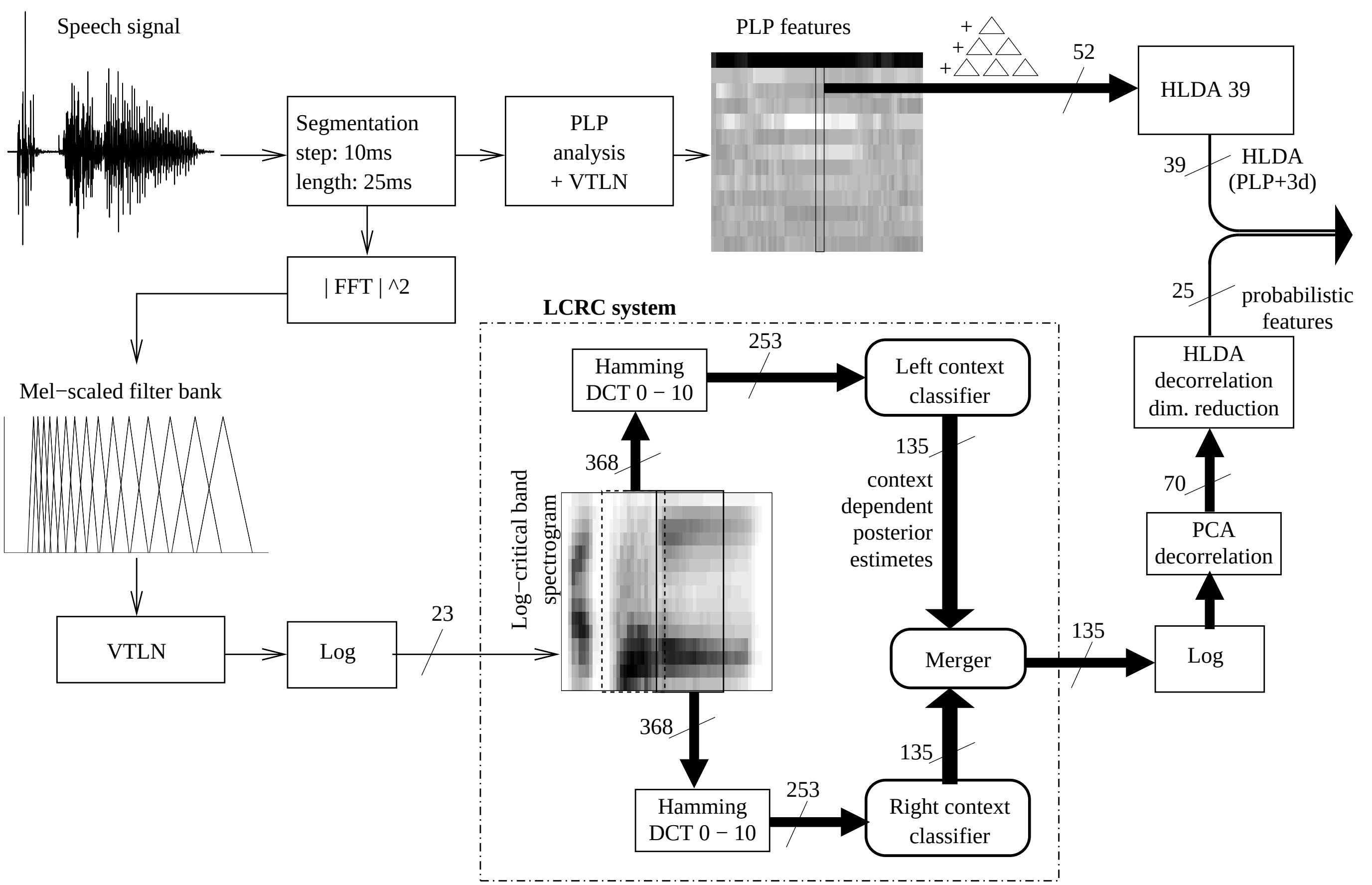
LM Components

LM component	word count	conference weight 2g	conference weight 3g	conference weight 4g	lecture weight 2g	lecture weight 3g	lecture weight 4g
AMI data from rt05s	206K	0.051	0.038	0.040			
CHIL rt06strain	76K				0.215	0.173	0.167
Fisher	21M	0.214	0.237	0.219	0.036	0.055	0.052
Hub4 LM96	151M	0.028	0.044	0.051			0.019
ICSI meeting corpus	0.9M	0.093	0.080	0.067	0.203	0.161	0.144
ISL meeting corpus	119K	0.126	0.091	0.091	0.023	0.020	0.017
NIST meeting corpus	157K	0.085	0.065	0.064			
Switchboard/callhome	3.4M	0.057	0.070	0.063		0.016	0.014
webdata (meetings)	128M	0.198	0.163	0.155	0.433	0.389	0.375
webdata (fisher)	128M	0.066	0.103	0.144			0.026
webdata (rt06s-conf)	138M	0.081	0.108	0.106		0.036	0.036
webdata (rt06s-lect)	114M				0.089	0.150	0.149
PPL RT06dev		109.2	88.1	84.5			
RT05S PPL		**106.9**	**86.2**	**82.7**	**157.9**	**127.6**	**122.4**
RT05S PPL - 2005 LM		105.6	84.3	81.2	165.6	137.4	134.5

Acoustic Modelling

- Same training data as in 2005 !
  - Both IHM and MDM
  - IHM 112 hours / MDM 65 hours !
  - IHM uses adaptation from 300hour CTS models
- Modelling basics
  - Decision tree state clustered triphones
  - CMN/CVN
  - MPE
  - HLDA
  - VTLN

Posterior features



- MLPs trained on 34 hours of speech

SAT

- Constrained MLLR (CMLLR) based SAT
- In addition to CMN/CVN and VTLN !

System	WER [%]
no adapt	28.7
adapt	27.9
1.SAT iter.	27.6
2.SAT iter.	27.4

System	WER [%]
no adapt	25.2
adapt	24.2
1.SAT iter.	24.1
2.SAT iter.	24.0
3.SAT iter.	23.9
4.SAT iter.	23.9

with posterior features

Results on RT05Seval

Discriminative Training

- Up to 15 iterations of MPE
- Word lattices generated with ML/PLP system

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

- Models trained with SAT and MPE on posterior features are denoted as ***M2 models*** later.

Alternative: Adaptation of CTS Models

- 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 ... $\Rightarrow$***M3 models***

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

Training on ihmtrain05, Results on RT05sEval

- MAP model adaptation from CTS

	CTS prior	CTS SAT prior
WB HLDA SAT system	27.4	27.4
1.SAT iter	26.7	26.9
2.SAT iter	-	26.5

Adaptation or training on ihmtrain05, results on RT05sEval

Including Discriminative Training

- Strategy

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

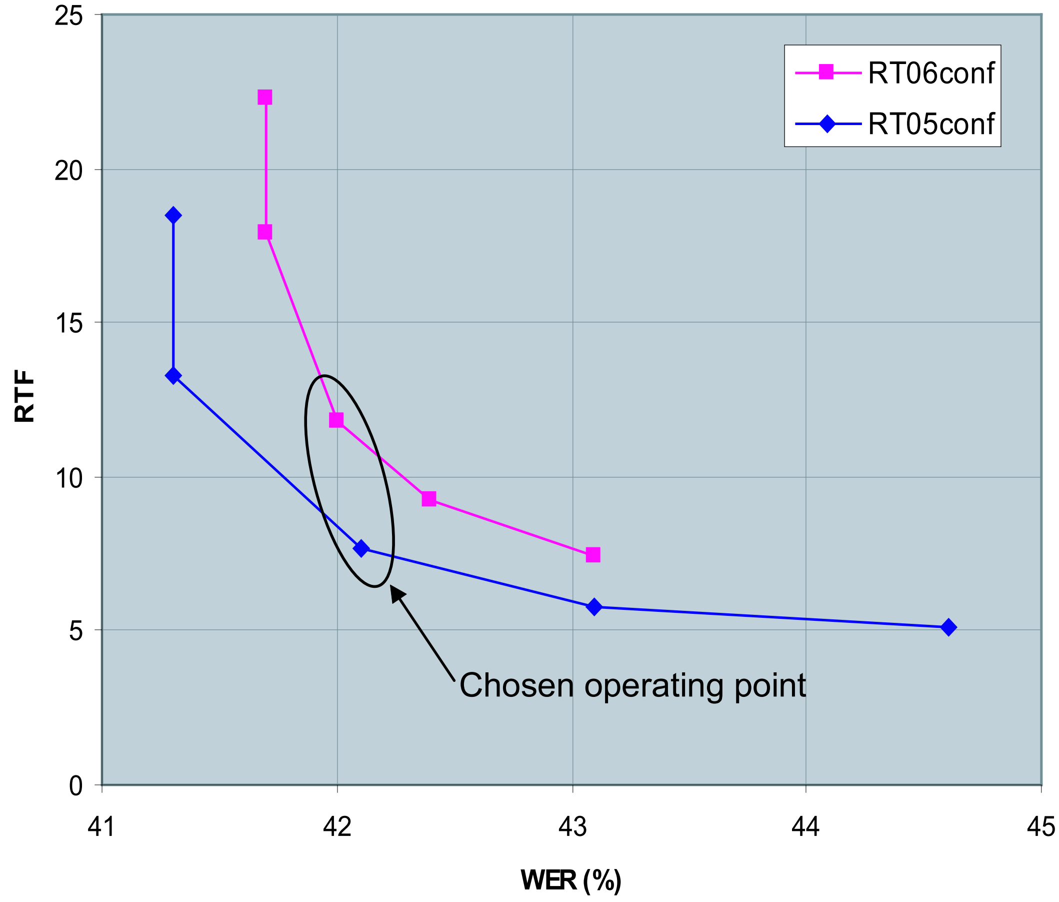
Initial models	Adaptation	WER [%]
CTS-SAT-MPE	-	30.4
CTS-SAT-MPE	ML-MAP	26.0
ML-MAP	MPE-MAP	23.9

Results on RT05sEval

Juicer - A WFST Decoder

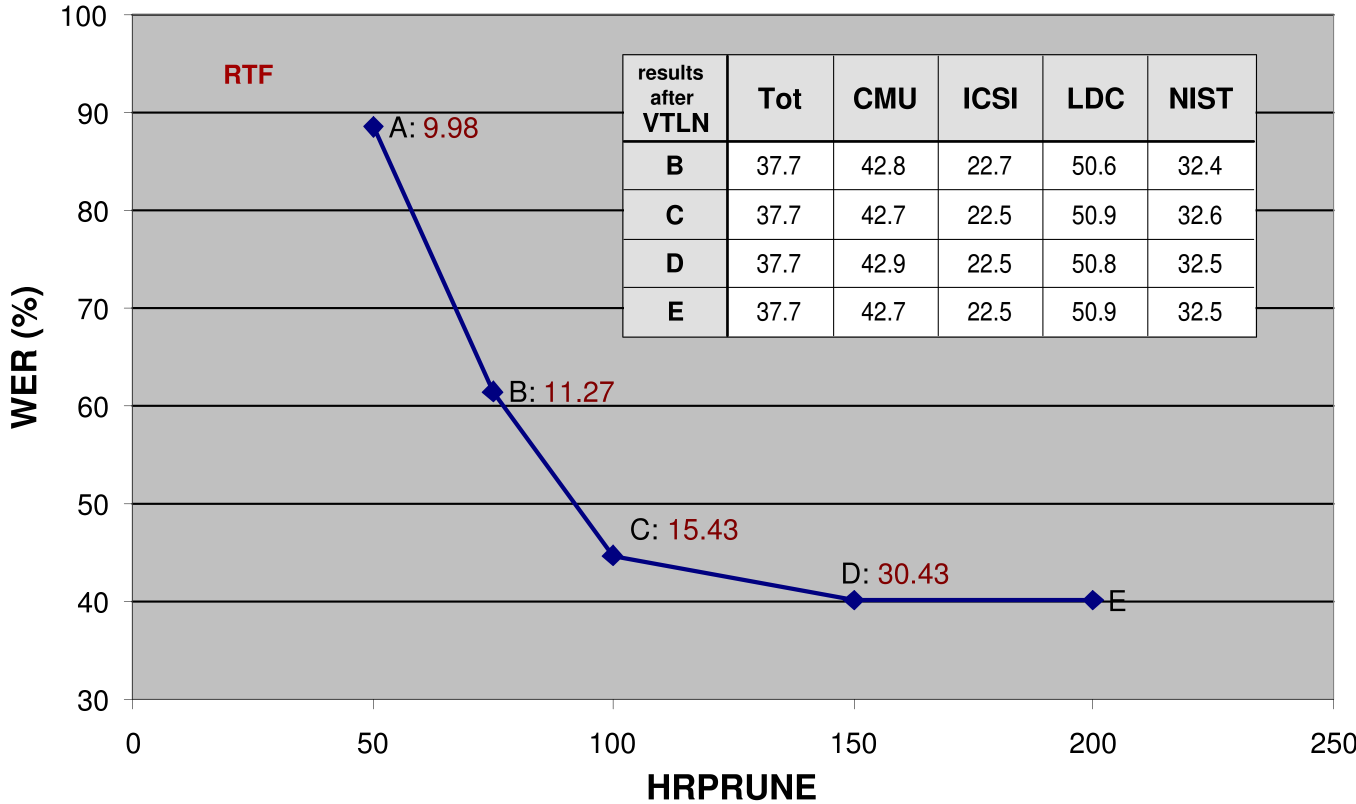
- A large vocabulary speech decoder based on weighted finite-state transducer (WFST)
  - Viterbi search with main-beam, model-end and histogram pruning
  - Static WFST composition using AT&T finite-state machine library and MIT FST toolkit
  - Favourable RTF vs WER when using tight pruning settings
- In development
  - Dynamic network composition
  - Lattice generation

Juicer – WER vs RTF



Speeding up VTLN

Performances of the 1st pass of decoding changing HRPRUNE and after VTLN on rt04seval IHM



results after VTLN	Tot	CMU	ICSI	LDC	NIST
B	37.7	42.8	22.7	50.6	32.4
C	37.7	42.7	22.5	50.9	32.6
D	37.7	42.9	22.5	50.8	32.5
E	37.7	42.7	22.5	50.9	32.5

warping factorsRMSE

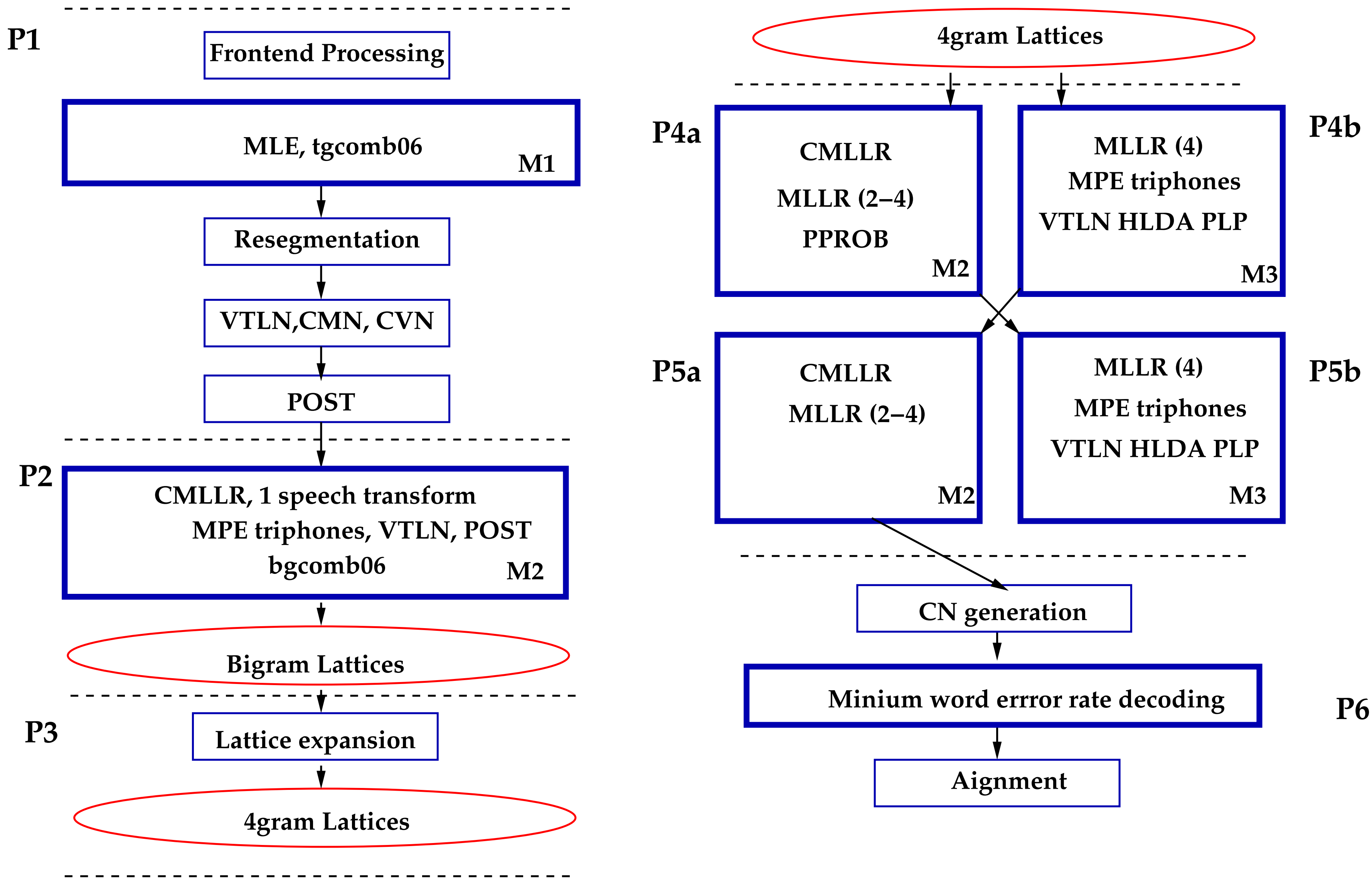
	B	C	D	E
B	-	-	-	-
C	0.0164	-	-	-
D	0.0162	0.0023	-	-
E	0.0164	0.0023	0.0001	-

Window-based MLLR

- MDM: addressing locally changing channels
- CMLLR transform estimated in a moving window
- Preliminary: no overlapping between windows

	TOT	Sub	Del	Ins	AMI	CMU	ICSI	NIST	VT
MLLR global	50.4	31.1	14.6	4.7	44.7	47.0	45.0	48.9	59.7
CMLLR global	50.5	31.3	14.5	4.7	44.6	48.7	44.8	50.4	58.7
CMLLR 1 min win.	50.3	31.0	14.5	4.8	44.1	49.9	44.9	49.0	58.9
CMLLR 2 min win.	50.0	30.6	14.7	4.7	44.8	47.9	44.6	48.4	58.5
CMLLR 5 min win.	50.0	30.9	14.4	4.7	44.0	47.7	45.2	49.1	58.6

System architecture



Results RT05 - Conference

- IHM

	TOT	Sub	Del	Ins	AMI	CMU	ICSI	NIST	VT
P1	37.9	22.8	11.2	4.0	38.5	35.8	30.9	44.0	41.2
P3.fg	25.4	13.5	9.5	2.5	24.6	21.8	22.6	31.8	26.7
P4a-cn	24.3	12.5	9.9	1.9	23.2	20.9	21.6	30.1	26.1
P5a-cn	23.7	12.0	9.9	1.7	22.0	20.1	21.1	30.0	25.7

- MDM

	TOT	Sub	Del	Ins	AMI	CMU	ICSI	NIST	VT
P1	52.4	33.3	14.5	4.6	49.5	52.5	50.7	53.1	55.2
P3.fg	35.4	20.8	11.5	3.1	31.7	34.0	38.0	38.4	35.7
P4a-cn	33.0	18.7	12.3	2.1	28.8	32.6	35.8	35.4	33.7

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
P2a	**29.2**	15.9	10.8	2.5	29.2	27.4	27.7	29.5	32.4
P3.tg	**26.6**	14.3	9.7	2.6	26.3	25.2	25.7	27.0	29.9
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
P5b-CN	**25.4**	13.1	10.2	2.1	25.2	23.5	24.8	24.2	29.8

MANUAL SEGMENTATION

	TOT	Sub	Del	Ins	CMU	EDI	NIST	TNO	VT
P1	40.3	27.0	8.5	4.9	40.4	39.5	38.7	37.6	40.9
P2a	26.5	17.3	6.8	2.5	26.7	25.5	26.6	22.3	28.8

Results RT06S - Conference - 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

Results RT06S - Lecture

- IHM

Pass	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
P2a	61.1	32.3	22.9	5.9
P3	59.3	31.6	21.2	6.5
P4a	58.7	29.2	23.9	5.7
P4a-cn	58.1	28.7	23.9	5.5

Conclusions/Summary

- Substantial improvement on both IHM and MDM
  - Substantially improved IHM front-end
  - Posterior features
  - Many smaller things
- Faster system
  - ~ 60x RT
- **THANKS**
  - All people in AMI for helping with getting our system together
  - ICSI/SRI for providing MDM segmentation and speaker information