IBM T. J. Watson Research Center

The IBM RT06S Speech-to-Text Evaluation Systems

Jing Huang, Martin Westphal, Stanley Chen

Olivier Siohan, Daniel Povey, Vit Libal, Alvaro Soneiro

Henrik Schulz, Thomas Ross, Gerasimos Potamianos

 © 2006 IBM Corporation

Outline

- Data for training and development
- Segmentation and speaker clustering
- Language and Acoustic Modeling
- Decoding Passes
- Results on the development data
- Results on the eval06 data
- Conclusions

 © 2006 IBM Corporation

Data for Training/Development

- **Training data:**
  - ICSI meeting (70 hours)
  - NIST meeting pilot (15 hours)
  - RT04 dev/eval (2.5 hours)
  - RT05 dev, excluding CHIL'05 eval (4.5 hours)
  - AMI seminars (16 hours)
  - CHIL06 dev (3 hours)
  - CHIL04 summer (6 hours, IHM data only)

  Total ***470-hour of MDM*** training data, and ***120-hour IHM*** training data

- **Development data:**
  - CHIL eval run #1 data, to measure our progress from last year
  - 1.8 hours of IHM data and 8.7 hours of MDM data

 © 2006 IBM Corporation

Segmentation and Speaker Clustering

- *Speech/non-speech segmentation*: long silence segments are discarded
- *Change-point detection* [J. Ajmera and C. Wooters, 2003]: speech segments are chopped into homogeneous regions
- *Segment clustering*: each segment is modeled by a single Gaussian; all Gaussians are clustered into a fixed number of clusters (say 4) using a Mahalanobis distance
- Same input features as those used in decoding

 © 2006 IBM Corporation

Manual segmentation vs. auto segmentation

- Auto. segmentation is actually better than manual segmentation. This may be due to the fact that transcribers tend to chop a whole sentence for easier processing.

WERs	Manual	Auto
MDM	55.4	53.5
IHM	35.6	32.4

SI decoding of dev data with manual/auto. segmentation

 © 2006 IBM Corporation

Language Modeling

Language Model

- Meeting transcripts (1.5M words), conference paper text (37M words) and Fisher data (3M words)
- The interpolated language model is pruned to about 5M n-grams (n<=4) and used to build a decoding graph
- The interpolation weights were 0.56, 0.37 and 0.07 respectively.
- LM rescoring uses the interpolated LM without pruning.

Lexicon

- 37K word, words in meeting and Fisher data, and 20K most frequent words in the other text corpus

NIST RT06S Workshop, May 3, 2006 © 2006 IBM Corporation

Old/New LM comparison

- Our last year's LM was built with CHIL transcripts (19K words) conference paper text (1.3M words) and Fisher data, with 2M n-gram (n<=3), and 20K lexicon

Data	Dev data	Dev data	chil06 eval
LM	old	new	new
perplexity	136.7	110.4	119.0
oov rate	3.9	0.4	1.0

 © 2006 IBM Corporation

Acoustic Modeling

Acoustic Features

- 13-dimensional PLP coefficients, LDA to 40-dim
- Mean normalized on a per speaker basis

Acoustic Models

- Quinphone statistics for decision trees
- 42 speech phone, 1 silence phone, 3 noise phones
- Speaker-Independent models:

  6K states/200K Gaussians for MDM data

  5K states/120K Gaussians for IHM data

 © 2006 IBM Corporation

SAT and fMPE

Three SAT models for MDM are built:

Model A: VTLN with variance normalization/SAT

Model B: VTLN no variance normalization/SAT

Model C: SAT after SI, no variance normalization

All MDM models have about 10K states/320K Gaussians,

the IHM SAT has 6K states/240K Gaussians, following ***Model A*** built.

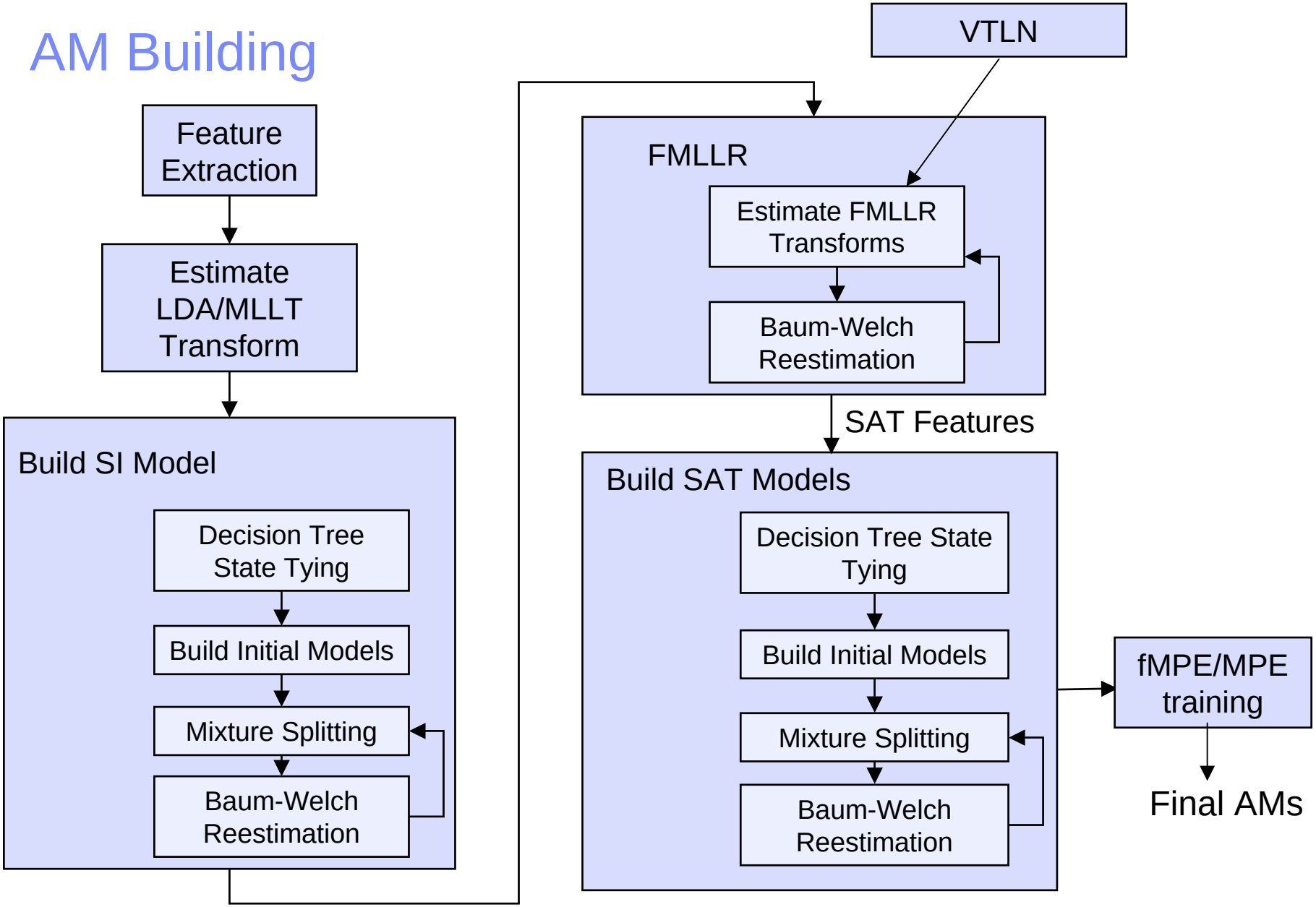
fMPE/MPE:

Transform computed from 1024 Gaussians obtained by clustering SAT models and projecting the posterior-based observation space to a 40-dimensional feature space

MAP-MPE: trained in the fMPE feature space with CHIL data as MAP adaptation data. MAP is necessary because of quite different acoustic conditions on CHIL data from the rest of meeting data.

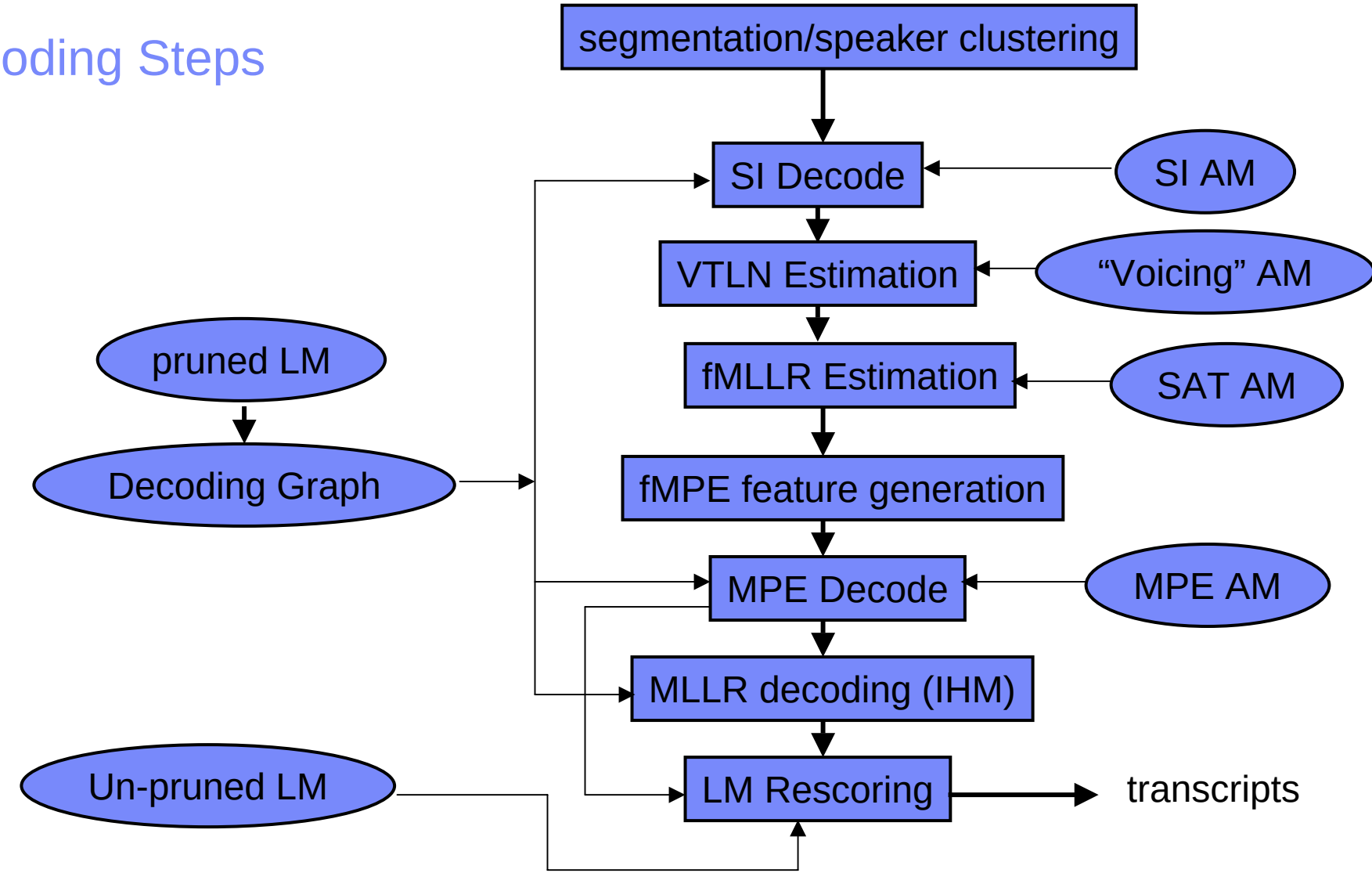
 © 2006 IBM Corporation

AM Building



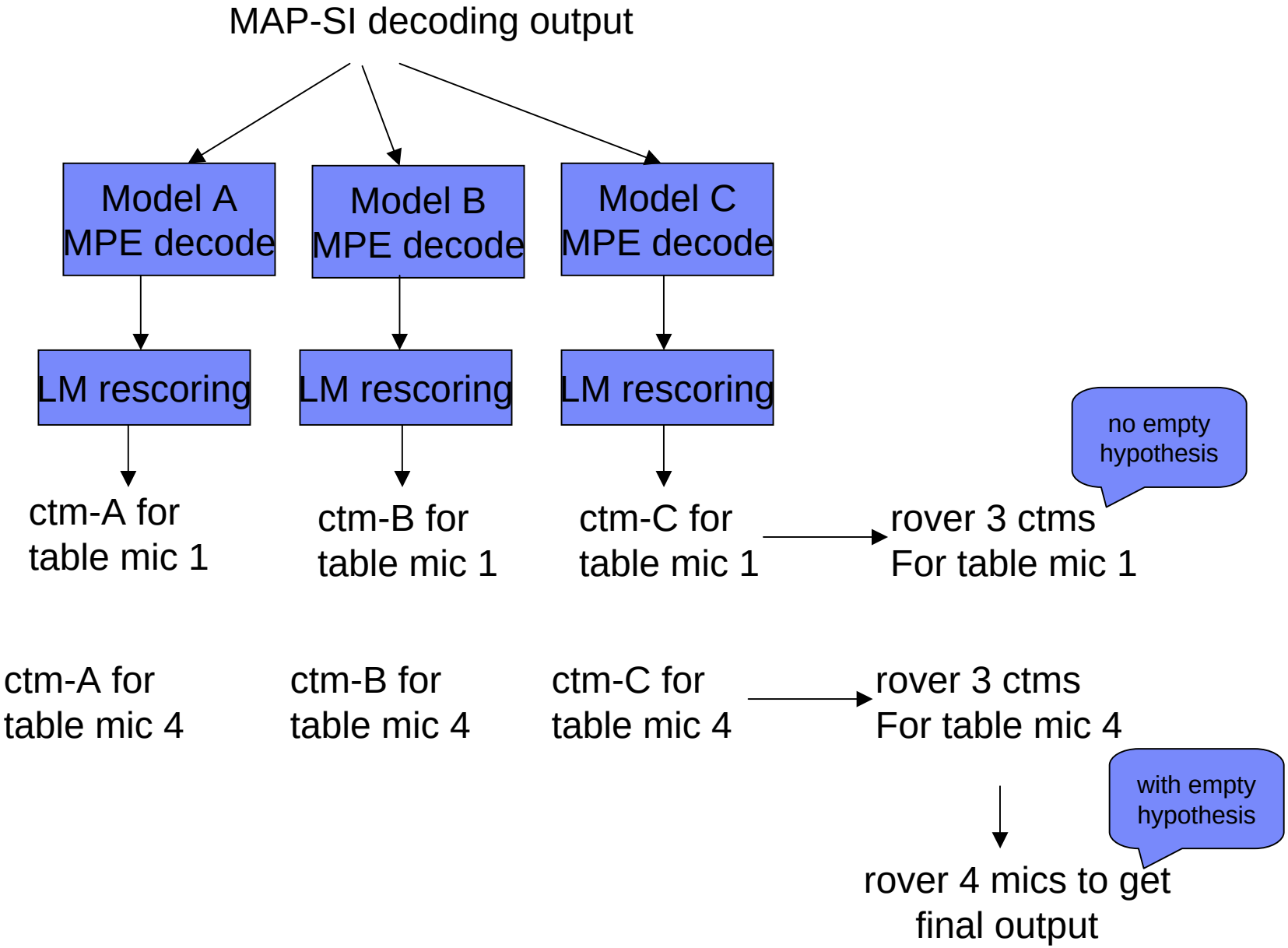
 © 2006 IBM Corporation

Decoding Steps



 © 2006 IBM Corporation

Rover



 © 2006 IBM Corporation

Results on IHM dev data: manual segmentation

Decoding Steps	Old LM	New LM
SI	39.8	35.6
MAP-SI	38.2	34.3
VTLN	38.1	-
SAT	36.9	-
fMPE	33.4	29.9
MAP-MPE	-	28.5
MLLR	-	26.8
LM rescoring	-	25.4

 © 2006 IBM Corporation

Results on IHM eval06 data (4/20 release)

Decoding Steps	Reference segmentation	Automatic segmentation
MAP-SI	42.9	72.5
fMPE/MAP-MPE	30.4	63.6
MLLR	29.3	63.5
LM rescoring	**28.3**	62.3

- Our cross-talk removal failed, degrades the WER from 62.3% to 66%

  after fixing bug, WER = 55.1%

- There is a huge gap between manual segmentation and automatic segmentation

 © 2006 IBM Corporation

Results on MDM dev data: manual segmentation

system	Model A	Model B	Model C
SI	55.4	55.4	55.4
MAP-SI	53.1	53.1	53.1
VTLN	55.0	53.8	-
SAT	53.4	51.4	52.1
fMPE/MAP-MPE	49.0	47.5	46.6
Final rover	-	-	45.6

 © 2006 IBM Corporation

LM rescoring on MDM dev data

	Model A	Model B	Model C
Table Mic	MPE/LM resc	MPE/LM resc	MPE/LM resc
1	53.5/51.3	52.0/49.6	51.3/48.7
2	55.0/52.7	54.1/52.0	53.4/51.1
3	55.1/52.7	53.8/51.7	53.5/51.7
4	55.3/53.5	54.4/51.9	53.9/51.7
rover table mics	49.0/49.0	47.5/46.8	46.6/47.8
Final rover	-	-	45.6/45.0

Remark: LM rescoring helps 2% absolute for each mic, normalizes output text, get little gain over MPE after rover

 © 2006 IBM Corporation

Results on MDM eval06 data (4/20 release)

system	Model A	Model B	Model C
MAP-SI	61.2	61.2	61.2
fMPE/MAP-MPE	51.8	51.6	52.5
LM rescoring	52.2	51.8	52.9
Final rover	-	-	51.1

- the above numbers are scored with o1 option
- IBM submission MDM output was not properly rovered, 52.0% --- so far the **best MDM and SDM** results
- Rovering by sequence of Model B, A, C, WER = 50.9%

 © 2006 IBM Corporation

Summary/Future Work

- Progress from 2005 on CHIL 05 eval: 36.9% → 25.4% on IHM data
- Simple yet successful system building for RT06 lecture meeting evaluation
- Rovering multiple table mics could give absolute gain of 4.5%

  about the same gain we see by using beamformed data
- Rovering multiple systems only give 0.5% absolute gain
- Better cross-talk removal for IHM data
- Better segmentation/speaker clustering
- Better frond-end with noise removal

 © 2006 IBM Corporation