IBM T. J. Watson Research Center

The IBM RT07 Speech-to-Text Evaluation Systems

Jing Huang, Etienne Marcheret,
Karthik Visweswariah, Vit Libal,
Gerasimos Potamianos

Outline

- Changes from last year
- Data for training and development
- Speaker Segmentation
- Language and Acoustic Modeling
- Decoding Passes
- Results on the development data
- Results on the eval07 data
- Conclusions

Changes from Last Year

- Improved SAD
- Improved SPKR
- ROVER of multiple systems built from randomized decision trees
- Huge language model decoding

Data for Training/Development

- **Training data:**
  - ICSI meeting (70 hours)
  - NIST meeting pilot (15 hours)
  - RT04 dev/eval (2.5 hours)
  - RT05 dev (6 hours)
  - AMI seminars (16 hours)
  - CHIL03/04 data (4 hours)
  - CHIL06/07 dev (6 hours)
  - 11 five-minute segments from CHIL RT06s eval

  Total 500-hour of MDM training data (did not re-train IHM AM)

- **Development data:**
  - 17 five-minute segments from CHIL RT06s eval

SAD

- HMM-based Speech/non-speech decoder
- Non-speech phones include silence and all three noise phones (different from last year)
- Choose an optimal operating point of false-alarm/miss by varying the number of Gaussians used for speech/non-speech

SPKR

- **Last year:** each segment is modeled by a single Gaussian in PLP space; all Gaussians are clustered into a fixed number of clusters (say 4) using a Mahalanobis distance
- **This year:** switch to MFCC 19-dim (no energy); first over-segment the data, then merge closest clusters, stop when reaching threshold.

Improvement on SPKR

DER/WER	SAD	SPKR	SI decoding
06	15.5	70.1	61.2
07	4.9	9.2	54.2

Effect of SAD/SPKR on SI decoding of dev data

Improvement on SPKR (cont.)

WER	Ref seg	Auto seg (07)	Auto seg (06)
SI	54.1	54.2	61.2
MPE	45.8	46.0	50.6
MLLR-MPE	43.5	44.1	---

Comparison of ref SPKR and auto SPKR on STT decoding

Results on SASTT (5/7 release)

data	SPKR DER	STT WER	SASTT WER
Dev (MDM)	9.2	41.9	44.1
Eval (MDM)	27.6	44.3	52.0
Eval (SDM)	27.4	47.9	55.4

- Best WERs for SASTT
- SPKR on eval07 degrades so much

Language Modeling

- CHIL meeting transcripts; other meeting transcripts; conference proceedings; Fisher data; **525M web data.**
- Interpolate 5 ngrams and prune to 5M ngram for static decoding graph.
- Prune only the web data LM and interpolate 5 ngram to obtain the final 152M ngram for a dynamic decoder.

Lexicon

- 37K words (word list from 5 text resources to have the best coverage on dev data)

Old/New LM comparison

- Our last year's LM was built on 4 text resources with no web data

Data	dev data	dev data	CHIL07 eval	CHIL07 eval
LM	06	07	06	07
perplexity	143.4	147.6	121.6	115.7
oov rate	1.4	0.4	1.4	1.1

Old/New LM comparison (cont.)

WER	LM(06)	LM(07)	LM(07) huge
SI	55.2	54.2	---
MLLR-MPE	---	46.7	43.5

Comparison of 06/07 LMs: decoding results on reference seg. of dev data.

Acoustic Modeling

Acoustic Features

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

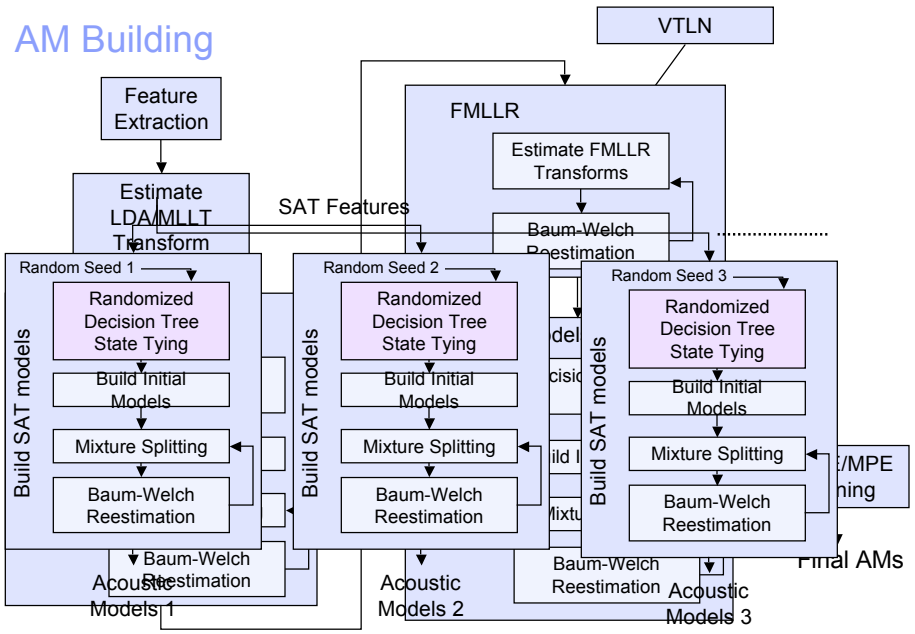
Acoustic Models

- Quinphone statistics for decision trees
- 42 speech phone, 1 silence phone, 3 noise phones
- Speaker-Independent models:
  6K states/200K Gaussians for MDM data

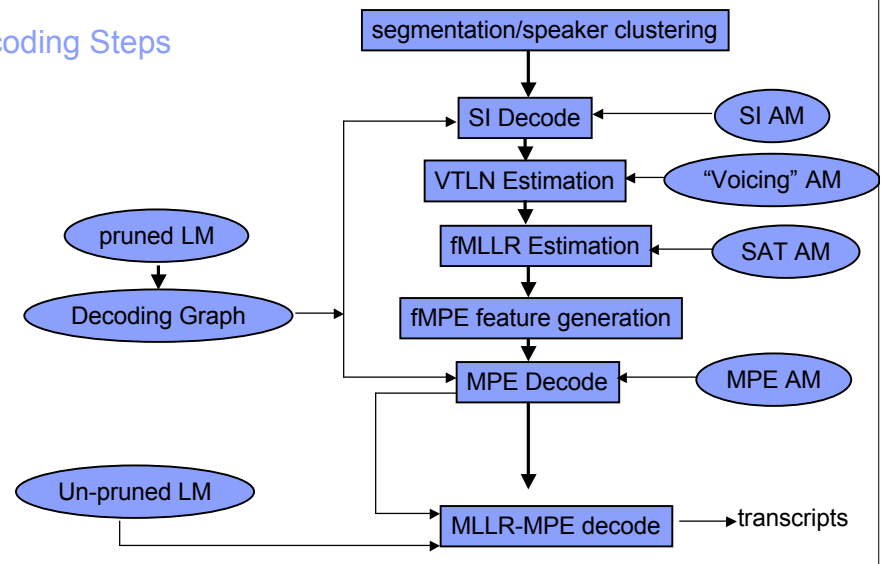
Building multiple systems

- **Use randomized decision tree:** a systematic approach to build multiple ASR systems
- **Systems only differ by their underlying HMM state tying**
  - The standard decision tree state-tying is modified to randomly select splits
- **Advantages**
  - All systems operate on the same feature set and use the same training data
  - The randomized decision tree state-tying is controlled by a single parameter, N (top-5 best candidate splits)

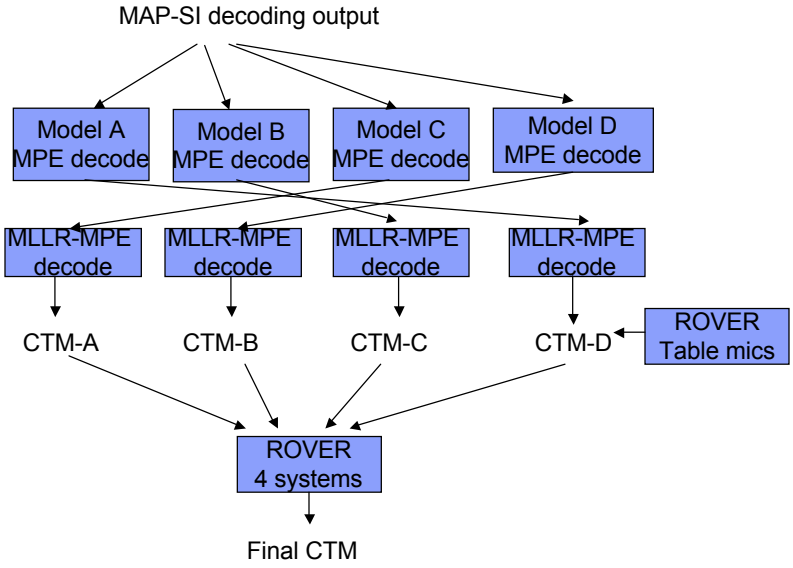
AM Building



Decoding Steps



Rover



Results on MDM dev data: with DER=9.2%

system	Model A	Model B	Model C	Model D
MAP-SI	54.2	54.2	54.2	54.2
MAP-MPE	46.3	46.0	47.3	46.3
MLLR-MPE	42.9	43.4	43.3	43.0
Final rover	41.9	---	---	---

Results on MDM eval07 data (5/7 release)

system	Model A	Model B	Model C	Model D
MAP-SI	55.5	55.5	55.5	55.5
MAP-MPE	48.9	48.6	48.6	48.9
MLLR-MPE	46.1	46.3	46.0	46.0
Final rover	44.3	---	---	---

- the above numbers are scored with o1 option
- Model A results were added late

Effect of Overlapping Speech (5/7 release)

Overlap factor	1	2	3
Final rover with empty hyp	44.3	48.3	50.0
Final rover with no empty hyp	45.0	48.7	50.3

- ROVER twice brings up high deletion rate, especially for overlapping speech!
  From O1 to O2, the deletion goes from 19.9% to 24.6%

Results on SDM eval07 data (5/7 release)

system	Model A	Model B	Model C	Model D
MAP-SI	58.6	58.6	58.6	58.6
MAP-MPE	53.7	53.1	53.2	53.4
MLLR-MPE	50.7	50.9	51.1	51.0
Final rover	47.9	---	---	---

- 3.6% worse than MDM results
- MDM gains from ROVER table mics

Results on IHM eval07 data (5/7 release)

Decoding Steps	Ref seg	Auto seg (SRI/ICSI)
MAP-SI	43.2	44.1
MAP-MPE	33.4	34.6
MLLR-MPE	31.7	33.4

- IHM AMs are the same as last year
- No multiple systems for ROVER

Summary

- Progress from part of eval06 data: 50.6% → 41.9%
- Rover of multiple table mics could give absolute gain of 3.6%
- Rove of multiple systems could get almost 3% absolute gain on SDM
- Need better speaker segmentation
- Explore noise robust front-end, beam-forming