Recent Developments at Cambridge in Broadcast News Transcription

D.Y. Kim, H.Y. Chan, G. Evermann, M.J.F. Gales,
D. Mrva, K.C. Sim & P.C. Woodland

Nov 9th 2004

Cambridge University Engineering Department

Presentation Overview

- RT03 Broadcast News System Review
- Training & Test Data
- Improved Acoustic Model Building
  - MPE training with MMI prior
  - Gender-dependent MPE training
  - NB model building using SPR-MPE
- RT04 Language Models
- RT04 Evaluation Systems
  - RT04 10xRT Primary System
  - RT04 10xRT Contrast System
  - RT04 1xRT System
- Post-Evaluation Experiments

RT03 CU-HTK BN-E Acoustic Models

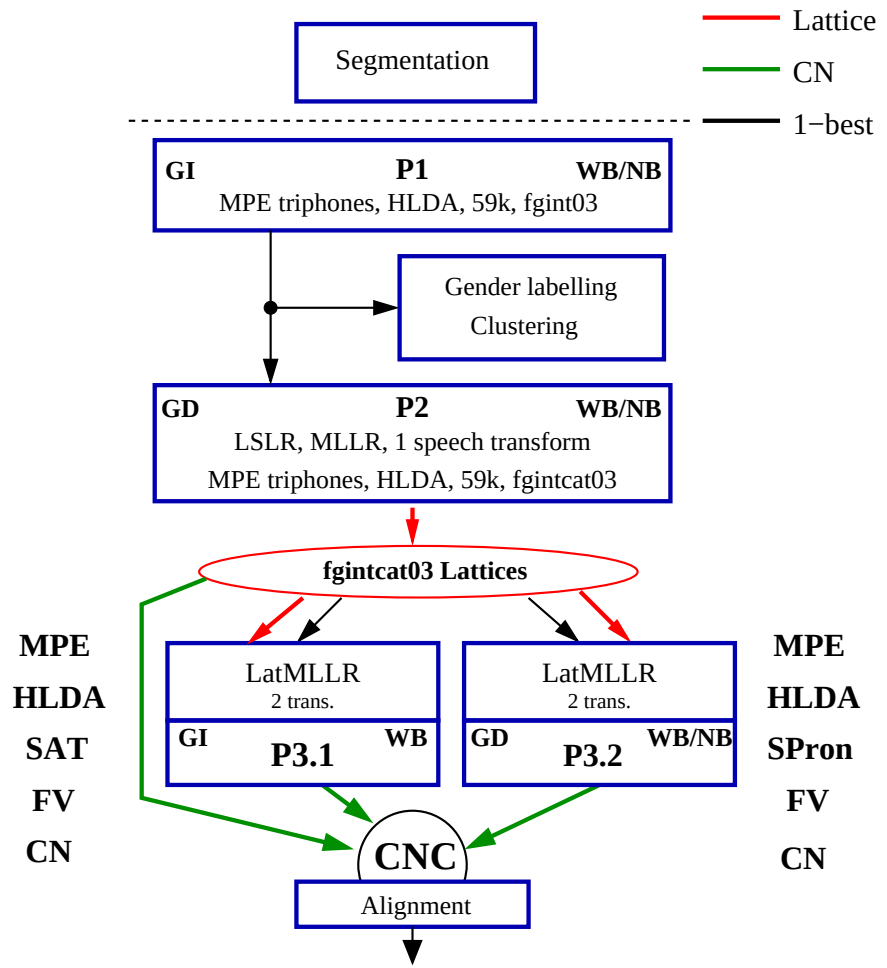
- Training data: the 144 hours acoustic BN training data from LDC
- Acoustic Models:
  - state-clustered, cross-word triphones
  - 7k tied states, 16 Gaussian components per state
  - HLDA projected 39-dim features
  - gender-dependent & bandwidth-dependent acoustic modelling
- Minimum Phone Error (MPE) training of all acoustic model
  - lattice re-generation & combination
  - MPE-MAP training for GD models
- SPron & SAT models for lattice re-scoring and system combination

RT03 CU-HTK BN-E Language Models

- 59k entry wordlist
- Word-based language models
  - training texts of 1 billion words in total
  - 5 subsets of training data
  - Good-Turing discounting with the HTK HLM toolkit for large subsets
  - modified Kneser-Ney discounting with SRI toolkit for small to mid-size subsets
  - entropy-based pruning after merging into a single model
  - pruned model has 8.8M bigrams, 12.7M trigrams, and 6.6M fourgrams
- Class-based language model
  - 1,000 automatically derived classes based on word bigram statistics

RT03 CU-HTK BN-E 10xRT System

- Segmentation
- Pass1: initial transcription
- Gender labelling / Clustering
- Pass2: lattice generation
- Pass3: lattice rescoring
  - P3.1: SAT
  - P3.2: SPron
- Confusion network combination
  - P3.1+P3.2+P2
- Ran in 9.1xRT on `eval03`



Acoustic Model Training Data for RT04

- Additional sources made available for RT04 model training

data	description	period	size(hours)
bnac	RT03 data	1996/97	144
tdt4	TV+radio	Oct00-Jan01	330
tdt4a	TV	Mar01-Jul01	530
bn03	TV	Mar03-Nov03	6375

Available audio data for BN task

  - tdt2 & tdt3 are also available (not used for RT04 system)
- For additional sources no manual transcriptions
- Available BN data dominated by TV shows (tdt4a & bn03)
  - radio shows from tdt4a period also released, no transcription/caption.

Lightly Supervised Training

Process to obtain training transcriptions:

1. Build a **biased language model** using available transcriptions
   - a data specific language model using closed caption text
   - interpolation of the data specific LM with a general LM
   - low perplexity for target data (hence biased)
2. Recognition with P1-P2 system
   - a simplified system architecture without lattice-rescoring
   - runs less than 5×RT
3. Post processing:
   - possible deletion of unreliable segments
   - tagging segments/words with confidence scores

Post Processing Experiments/Comparisons

- Investigated techniques for reliability of segment hypothesis:
  - selection based on closed-caption filtering
  - selection based on confidence-based filtering
  - modified MPE training for:
    - word/segment confidence scores in numerator lattices
    - both recognised/CC word alternatives in numerator lattices
- Compared training on `tdt4` corrected captions and lightly supervised
  - only 7% disagreement in word tokens between two transcripts
  - no significant difference in performance
- None of the approaches made a significant difference
  - use standard lightly supervised training with no selection/filtering

Training Data

- Four training sets used for development:

training set	description	size
bntr04-base	bnac+tdt4	375
bntr04-750h	+tdt4a	752
bntr04-1050h	+bn03_1	1050
bntr04-1350h	+bn03_2	1350

Selected BN-E training data sets and sizes

- Lightly supervised training for tdt4 & tdt4a
- Two 300hour subsets from BBN's 2515hour of bn03 transcriptions
  - bn03_1 300hrs from ABC, CNBC, CNN, CNNHL, CSPAN, PBS
  - bn03_2 300hrs from CBS, CNN, FOX, MSN, MSNBC, NBC, NWI

Test Data

- 4 sets of data were used for development

Test set	# Shows	Size	Period
dev03	6	3hrs	Jan01
eval03	6	3hrs	Feb01
dev04	6	3hrs	Jan01
dev04f	6	3hrs	Nov03

BN-E test sets and sizes

- dev04 shows selected by STT sites
  - dev03 and dev04 have 2 shows duplicated
- dev04f representative of the extended broadcast news corpus
- No epoch overlap with the acoustic training data.

Dynamic MMI Prior

- I-smoothing required for good generalisation of MPE:
  - standard scheme uses a *dynamic ML prior*
  - investigate IBM-style *dynamic MMI prior*
  - use *static GI-MPE prior* for GD models.

	dev03	eval03
MPE (dynamic ML prior)	13.9	12.6
+GD MPE-MAP	13.7	12.4
MPE (dynamic MMI prior)	13.6	12.5
+GD GI-MPE prior	13.5	12.3

Models built using `bntr04-base`. 16 comp/state. Single pass decoding with the RT03 trigram LM. NB segments decoded using the `RT03` MPE NB models.

- Consistent (small) gains with dynamic MMI prior;
- Consistent (small) gains with static MPE prior for GD modelling

Efficient Way to Build Narrow Band Model

- Small consistent gains from using band-dependent models (NB models)
  - computationally expensive to rebuild using ML SPR and MPE training
- MPE Single Pass Re-training (SPR) from MPE trained WB model-set
  - assume numerator and denominator "occupancies" similar for NB and WB
  - use NB ML statistics to get "current" model parameters

Training Method	Iter	%WER		
		dev03	eval03	dev04
NB MPE	8	14.9	13.6	16.5
MPE-SPR (ML prior)	–	15.0	13.8	16.6
+MPE	1	14.7	13.7	16.4

%WER with various bnac NB acoustic models. Single pass decoding with RT03 trigram LM. WB segments hypothesis using the RT03 WB MPE model.

- Similar performance using MPE-SPR to rebuilding using ML-SPR and MPE.

Increased Training Data/Model Complexity

- Investigate effects of increasing quantity of training data & components/states

Training Data		%WER			
		dev03	eval03	dev04	dev04f
bntr04-base	16/7k	13.6	12.5	–	–
bntr04-750h	16/7k	13.4	12.1	–	–
bntr04-750h	32/7k	12.8	11.8	13.8	21.6
bntr04-1050h	32/9k	12.2	11.4	13.1	20.3
bntr04-1350h	32/9k	12.1	11.2	13.2	19.6

Single pass GI MPE decoding of WB segments with the RT03 trigram LM. NB segments decoded using the RT03 NB MPE model.

- bntr04-base to bntr04-750h gave significant gains
- Increasing components/states gave additional gains
- Largest gains on dev04f by adding bn03 (closer epochs)

P1-P2 System Performance

Training Data		%WER			
		dev03	eval03	dev04	dev04f
bntr04-base	16/7k	11.6	10.7	13.3	20.0
bntr04-750h	16/7k	11.2	10.5	13.0	19.6
bntr04-750h	32/7k	10.9	10.2	12.8	18.9
bntr04-1050h	32/9k	10.5	9.7	12.2	17.6

%WER of the P1-P2 system with the RT03 LM. NB segments decoded using the RT03 NB MPE model.

- Additional training data and increased number of model parameters are still giving gains after adaptation

Language Model Training Corpus

Training text	Size(MW)
PSM's broadcast news transcripts 1992-99, TDT2&3 closed captions, LDC's broadcast news closed captions 2003	334
transcripts from CNN's website 1999-2000, 2001-2003	147
TDT4 closed captions 2000-01, TDT4a in 2001	5
NIST's broadcast news training data from 1997/98, Marketplace show transcripts	2
Newswire texts from Los Angeles Times and Washington Post 1995-98, New York Times 1997-2000 & 2001-2002, Associated Press 1997-2000 & 2001-2002	928

- Increased text corpus
  - 1.4 billion words in training (1 billion words in RT03)

Language Model Performance

- New word list, still 59k entries: reduced OOV rates in dev sets

	eval03	dev04	dev04f
RT03 wlist	0.66	0.57	0.54
RT04 wlist	0.45	0.49	0.42

- Pruned LM has 17M bigrams, 28M trigrams, and 23M 4-grams
- PPs for eval03, dev04 and dev04f were 120, 118, and 132.
- WER reductions of 0.3-0.5% abs with the new LM in P1-P2 framework.

LM	eval03	dev04
RT03	9.7	12.2
RT04	9.2	11.9

%WER in P1-P2 system with bntr04-1050h models.

CUED RT03 segments.

Improved/Dual Segmentations

- LIMSI 2003 segmenter used along with CUED segmenter
  - able to compare effects of two segmentations
  - examine effects of poor/failed segmentation

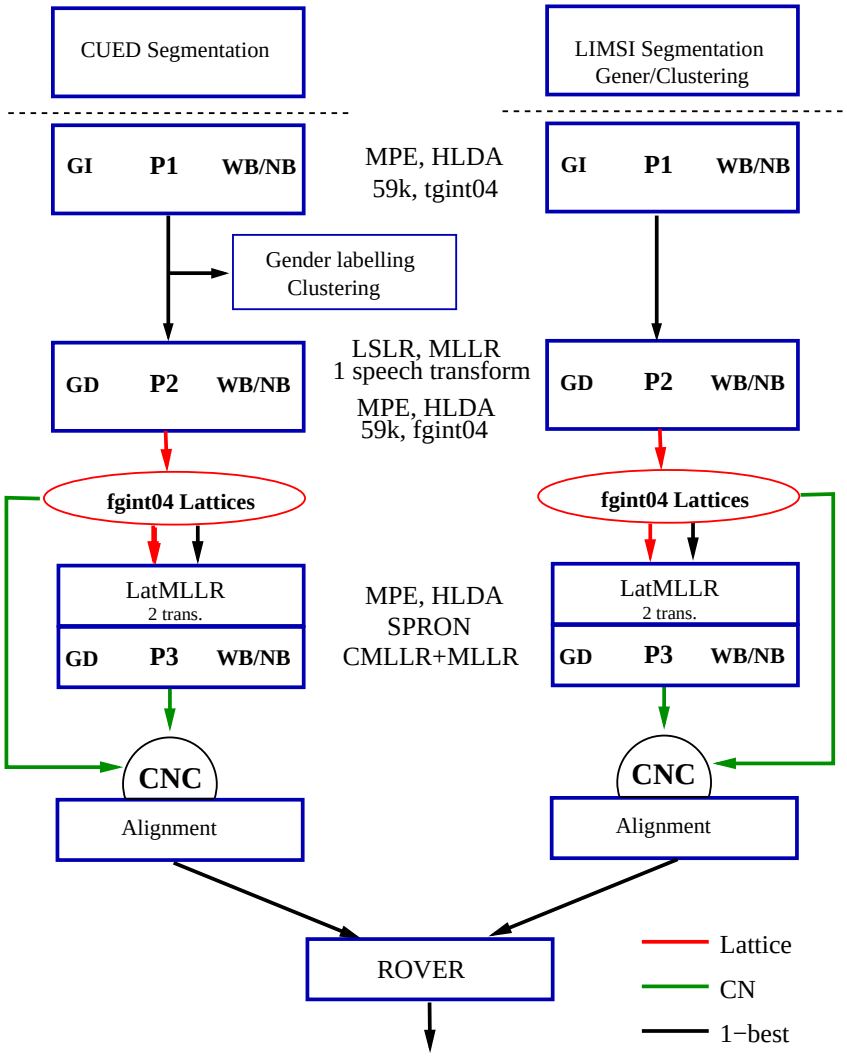
Segment	%WER		
	eval03	dev04	dev04f
CUED	9.2	11.9	16.6
LIMSI	8.8	11.4	16.2
ROVER	8.5	11.0	15.8

%WER of P1-P2 system and ROVER using CUED and LIMSI segmentations. bntr04-1050h WB models, the RT03 NB models. RT04 LM.

- LIMSI segmenter consistently better than CUED segmenter, 0.4% abs
- ROVER two segmentation outputs gave consistent 0.3-0.4% abs gain

BN-E RT04F 10xRT Primary System

- Two separate sub-systems:
  - sub-system 1: CUED segmenter
  - sub-system 2: LIMSI segmenter
- Each sub-system:
  - fast MPron P1 (no fg expansion)
  - P2: MPron `bntr04-1350h`, 3xRT
  - P3: SPron `bntr04-1350h`
  - CNC using P2 and P3
- Combining outputs using ROVER
- Ran in 9.9×RT on `eval04`



10×RT Primary System Results

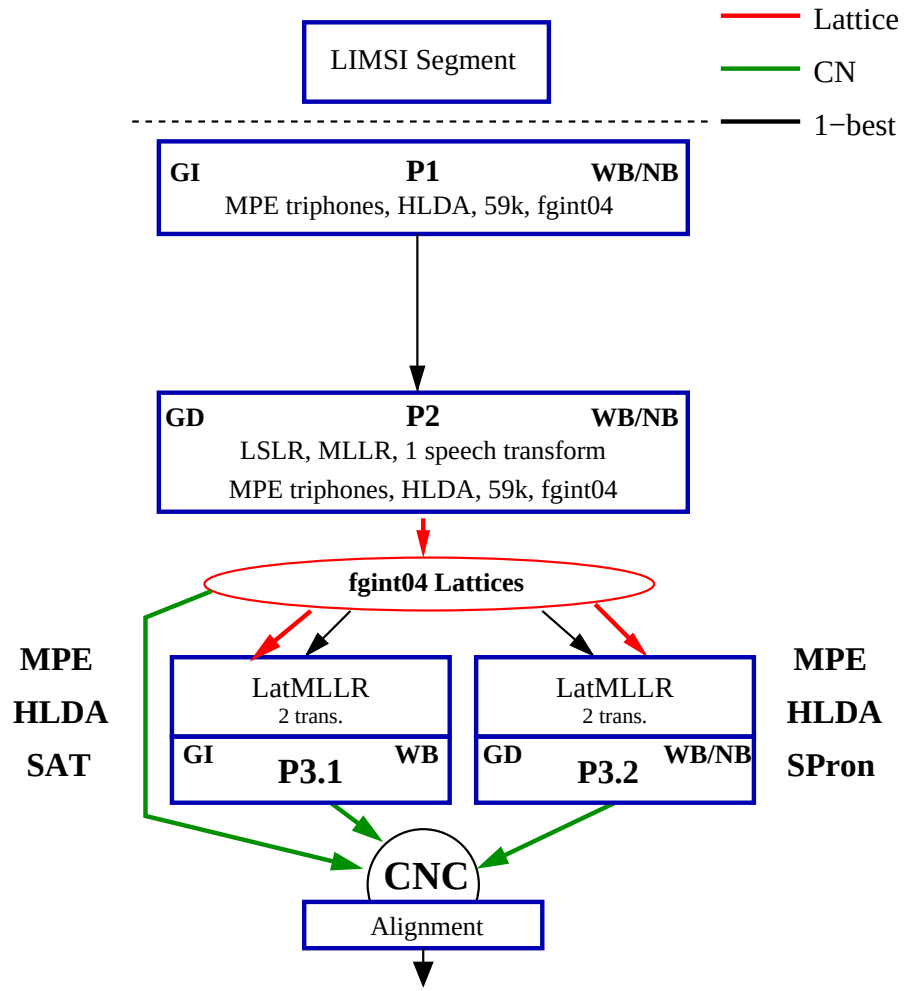
shows	CUED	LIMSI	ROVER
20031204_130035_cnn	**14.9**	**12.1**	**12.8**
20031203_183814_abc	17.5	16.4	16.3
20031217_184122_abc	16.6	15.8	15.7
20031215_204057_cnnhl	12.0	11.5	11.2
20031215_231058_wbn	11.5	11.0	10.8
20031218_004126_pbs	18.9	18.6	18.6
20031202_203013_cnbc	11.2	11.1	10.5
20031209_193152_abc	7.8	7.7	7.2
20031202_050216_cnn	10.2	10.2	9.9
20031209_193946_pbs	10.0	10.0	9.7
20031206_163852_cspan	**17.1**	**17.2**	**16.1**
20031219_202502_cnbc	**10.0**	**10.2**	**9.8**
Total	**13.3**	**12.8**	**12.6**

%WER and run-time of the RT04 10×RT primary systems on `eval04`

- Similar performance to development numbers (fairly consistent on show basis)

BN-E RT04F 10×RT Contrast System

- LIMSI Segmenter
- Similar structure as RT03S 10xRT
- Two P3 branches:
  - P3.1: SAT `bntr04-1050h`
  - P3.2: SPron `bntr04-1350h`
- System Combination
  - P3.1+P3.2+P2



10×RT Contrast Performance

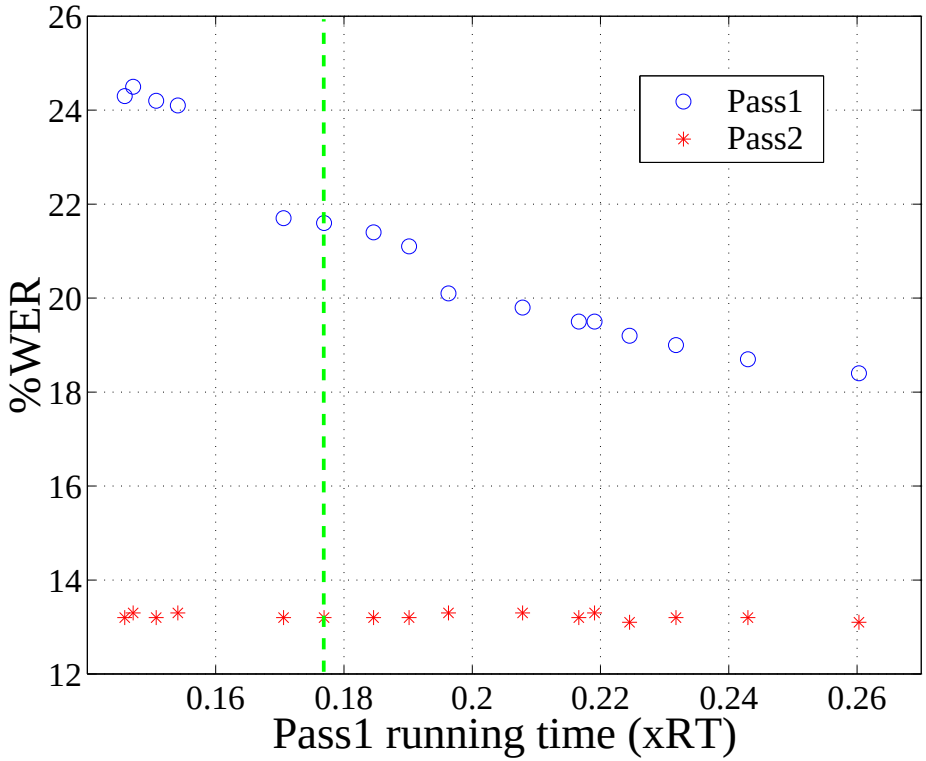
System		%WER		
		eval03	dev04	dev04f
RT03 10×		10.6	13.2	18.6
RT04 10× Contrast	P1	10.9	13.8	19.1
	P2	8.6	11.1	15.9
	P3.1	8.3	10.8	15.6
	P3.2	8.1	10.4	15.2
	Final	8.0	10.4	14.9

Performance of the Contrast system in comparison with the RT03 10×RT system.

- Consistent gains over 2003 RT03S system:
  - a 22% relative reduction in WER for dev sets
- small gains from confusion network combination
- Ran in 8.4×RT on eval04

CUED 1xRT System Design

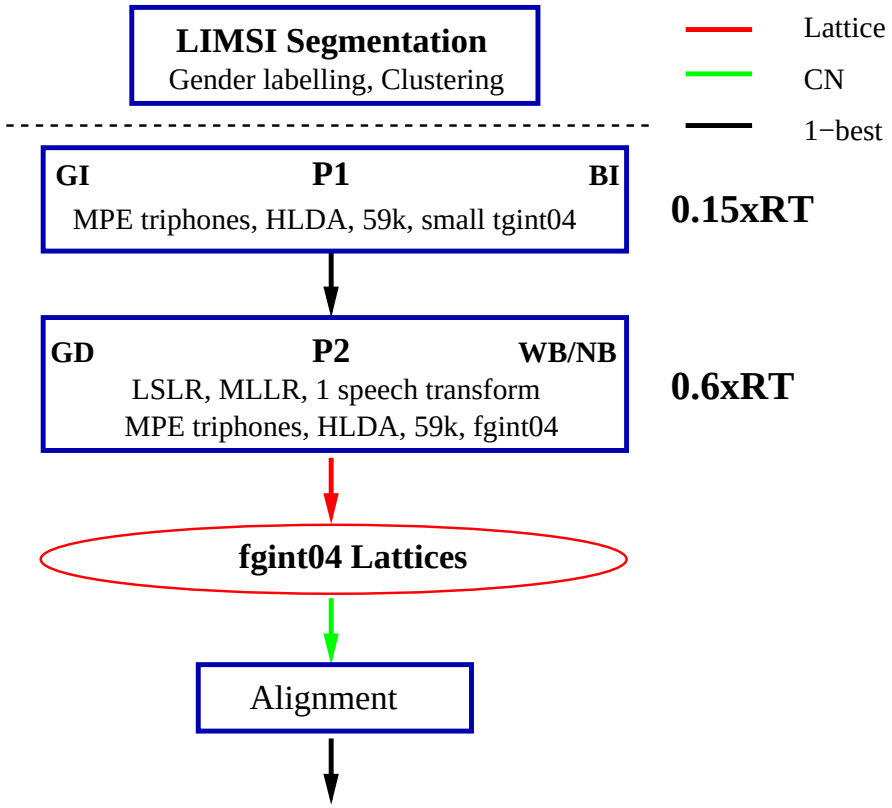
- Need to do adaptation estimation - fast initial P1 required



- Plot shows effect of P1 pass (in terms of xRT) on accuracy dev04
  - P2 relatively insensitive to P1 pass

CU-HTK RT04 1xRT System Structure

- LIMSI segmentation
- Very fast P1 system (0.15xRT)
- P2 pass (0.6xRT)
  - MPron `bntr-1350` trained
  - LSLR mean adaptation
  - diagonal variance adaptation
- Confusion network decoding
- Delete low confident words
- Forced alignment



CU-HTK RT04 1xRT System Performance

Pass	%WER			
	eval03	dev04	dev04f	eval04
P1	17.2	21.7	27.8	25.6
P2	9.9	12.7	17.4	15.4
Final	9.8	12.5	17.3	15.3

%WER of the RT04 1xRT system

- Only 21% worse on eval04 than the primary 10× system (12.6%)
- Better performance than the RT03 10× system

Performance Summary in RT03 & RT04

System		%WER		
		eval03	eval04	progress
10×	RT03	10.6	–	12.7
	RT04 Contrast	8.0	12.9	9.8
	RT04 Primary	7.8	12.6	9.4
1×	RT03	14.6	–	16.8
	RT04	9.8	15.3	11.8

System performance comparison in the RT03 and RT04 evaluations.

- 10×RT: 26% relative error reduction on progress set
- 1×RT: 30% relative error reduction on progress set

Post Evaluation: SAT and SPron-SAT

- SAT model re-training: using `bntr04-1350h` training set
  - improved branch performance, no difference after CNC
- SPron-SAT model
  - comparable performance with the SPron model

Pass		%WER		
		`eval03`	`dev04`	`dev04f`
P3.1-cn	SAT (1050h)	8.3	10.8	15.6
P3.1a-cn	SAT (1350h)	8.2	10.6	15.3
P3.2-cn	SPron	8.1	10.4	15.2
P3.3-cn	SPron-SAT	8.1	10.5	15.0
P2+P3.1+P3.2		8.0	10.4	14.9
P2+P3.1a+P3.2		8.0	10.4	14.9
P2+P3.2+P3.3		8.0	10.3	14.8

%WER of various P3 branches after confusion network decoding in the RT04 10xRT contrast system framework.

Post Evaluation: System Combinations

- More system combinations with various models
  - evaluated different acoustic models in the 10×RT primary system framework
  - a small gain on `eval04` with SPron & SPron-SAT combination

CUED-seg	LIMSI-seg	`dev04`	`eval04`
SPron	SPron	10.0	12.6
SPron	SPron-SAT	10.0	12.5
SPron-SAT	SPron	10.0	12.6
SPron-SAT	SPron-SAT	10.1	12.5

%WER for `dev04` & `eval04` using SPron & SPron-SAT models in the RT04 10×RT primary system.

Conclusions

- For the 10×RT system, a good relative gain of 26% was made on progress set based on
  - huge amount of training data with lightly supervised training
  - improvements in acoustic model training
  - increased language model
  - combining dual segmentations
- A high performing 1×RT system was developed which is better than the RT03 10×RT system