Recent Developments at Cambridge in Broadcast News Transcription

Do Yeong Kim, Mark Gales, Phil Woodland &
Rest of the CU-HTK STT Team

November 16th 2004

Cambridge University Engineering Department

Overview

- EARS & Broadcast News Transcription
- RT03 10xRT System
- Training & Test Data
- Improved Acoustic Model Building
  - MMI prior & Gender-dependent MPE training
  - MPE Single Pass Re-training
  - Effects of Increased Training Data
- Language Model Development
- RT04 Evaluation Systems
  - 10xRT Systems
  - 1xRT System
- SupearEARS: Cross-site BN System

EARS project & Broadcast News Transcription

- DARPA EARS programme
  - Speech-to-Text (STT) & Metadata
  - Broadcast News (BN) & Conversational Telephone Speech (CTS)
  - English, Madarin & Arabic

- Broadcast News English Transcription
  - US TV & radio broadcast news
  - difficulties due to heterogeneous data
    * many speakers including non-native speakers
    * various speaking styles: read/spontaneous/conversational
    * different channel conditions: wideband/narrowband
    * background music/noise

RT03 CU-HTK BN-E Acoustic Models

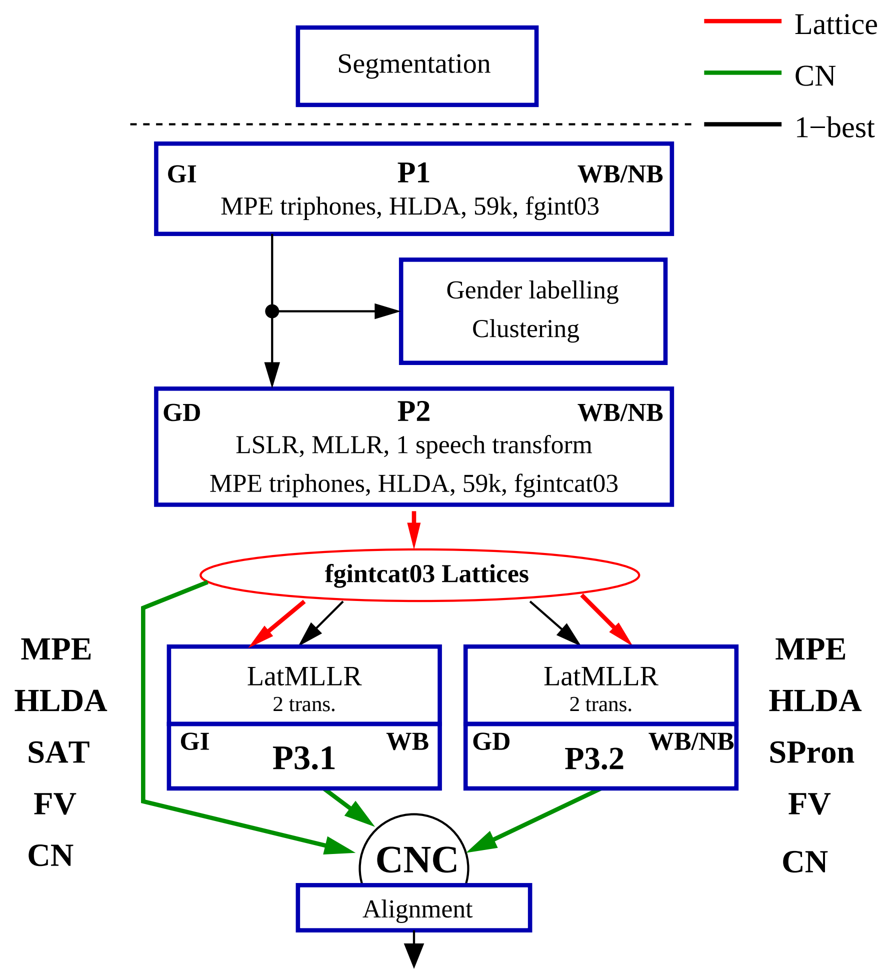
- Porting technologies from CTS to BN
- Training data: the 144 hours acoustic BN training data from LDC
- Acoustic Models:
  - state-clustered, cross-word triphones
  - 7k tied states, 16 Gaussian components per state
  - HLDA projected 39-dim features
  - gender-dependent & bandwidth-dependent acoustic 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
  - texts corpus of 1 billion words in total
  - training 4-gram language models on 5 subsets using HTK HLM toolkit & SRI toolkit
  - merging into a single model using interpolation weights optimised on dev data
  - after pruning the model has 8.8M bigrams, 12.7M trigrams, and 6.6M 4-grams
- Class-based language model
  - 1,000 automatically derived classes based on word bigram statistics
  - interpolated with the word-based language model

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
- 10.7% WER in 9.1xRT on `eval03`



Available Data for Acoustic Model Training

- Available audio data for BN task

	data	description			size(hours)
√	bnac	TV+radio	transcribed	1996/97	144
	tdt2	TV+radio	caption	Feb98-Jun98	640
	tdt3	TV+radio	caption	Oct98-Dec98	475
√	tdt4	TV+radio	caption	Oct00-Jan01	330
√	tdt4a	TV	caption	Mar01-Jul01	530
	tdt4a	radio	—	Mar01-Jul01	340
√	bn03	TV	caption	Mar03-Nov03	6375

- Huge amount of audio data with *no* manual transcription
  - closed captions available for TV shows

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
   - advertisement removal before segmentation
   - 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

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 extended broadcast news data
- No epoch overlap with the acoustic training data.

Dynamic MMI Prior

- I-smoothing required for good generalisation of MPE:

$$\mu_{jm} = \frac{\{\theta^{\text{num}}_{jm}(\mathcal{O}) - \theta^{\text{den}}_{jm}(\mathcal{O})\} + D_{jm}\hat{\mu}_{jm} + \tau^I \mu^{\text{ml}}_{jm}}{\{\gamma^{\text{num}}_{jm} - \gamma^{\text{den}}_{jm}\} + D_{jm} + \tau^I}$$

  - standard scheme uses a *dynamic ML prior*, μ^{ml}_{jm}
  - investigate IBM-style *dynamic MMI prior*, μ^{mmi}_{jm}
  - 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.

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		ML	MPE			
		eval03	dev03	eval03	dev04	dev04f
bntr04-base	16/7k	14.8	13.6	12.5	–	–
bntr04-750h	16/7k	14.6	13.4	12.1	–	–
bntr04-750h	32/7k	14.0	12.8	11.8	13.8	21.6
bntr04-1050h	32/9k	13.8	12.2	11.4	13.1	20.3
bntr04-1350h	32/9k	13.6	12.1	11.2	13.2	19.6

%WER of single pass GI decoding of WB segments with the RT03 trigram LM. NB segments decoded using the RT03 NB models.

- 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 LMs. 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

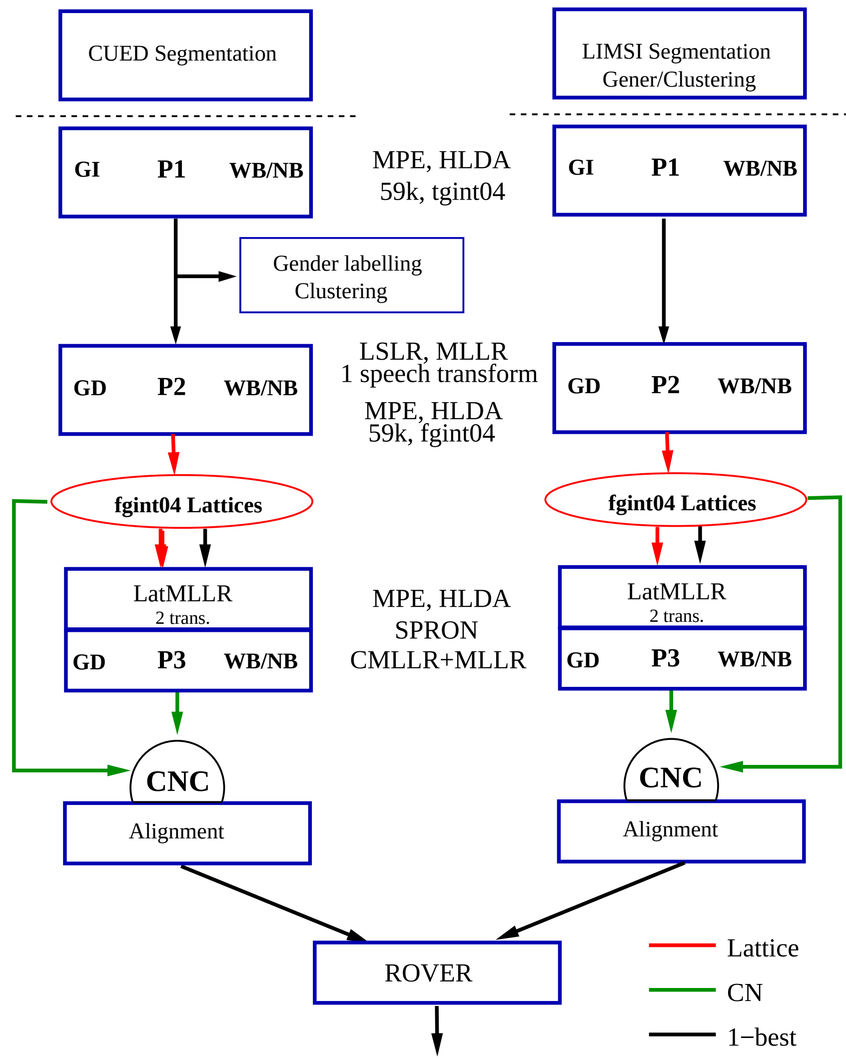
Segment	%WER		
	eval03	dev04	dev04f
CUED	9.2	11.9	16.6
LIMSI	8.8	11.4	16.2
ROVER	8.5	11.0	15.8

%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

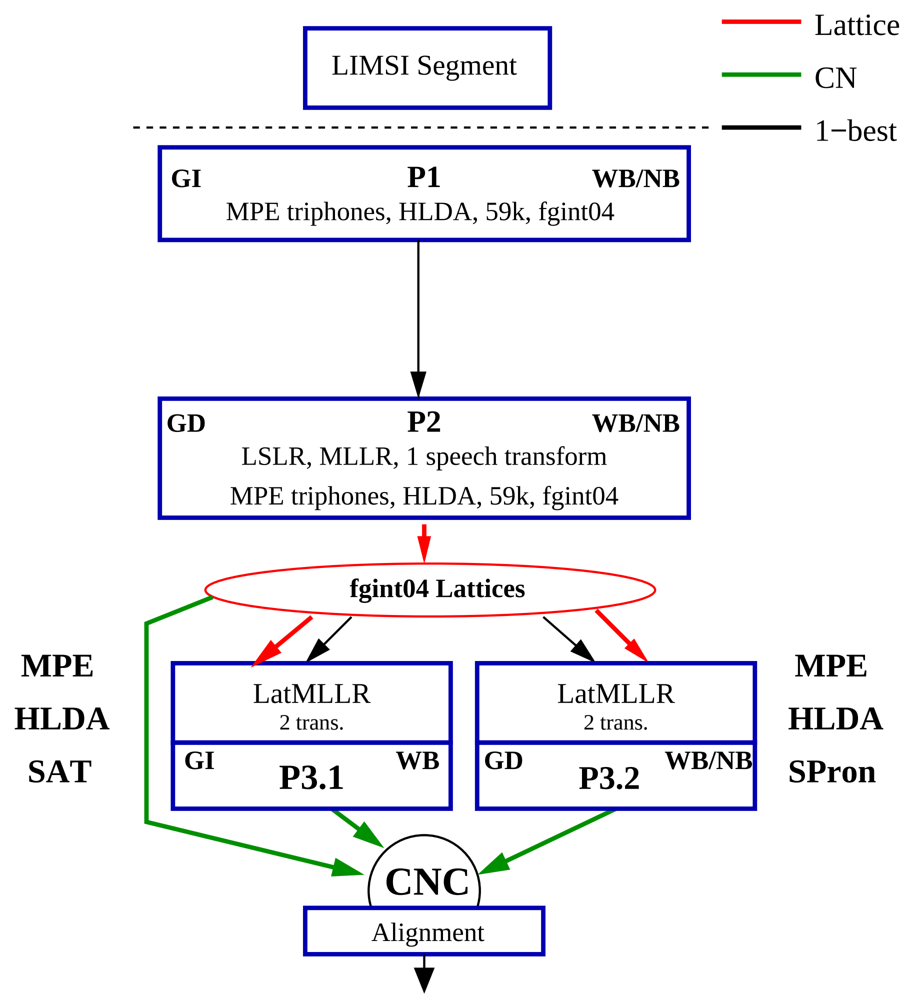
BN-E RT04F 10xRT Primary System

- 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`



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

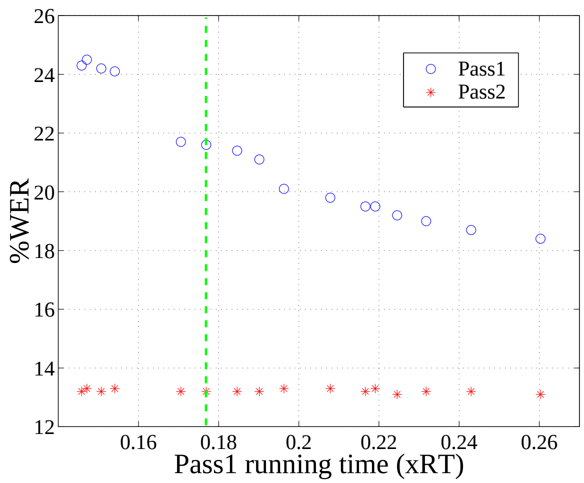
System		%WER		
		eval03	dev04	dev04f
RT03 10×		10.6	13.2	18.6
RT04 10× Contrast	P1	10.9	13.8	19.1
	P2	8.6	11.1	15.9
	P3.1	8.3	10.8	15.6
	P3.2	8.1	10.4	15.2
	Final	8.0	10.4	14.9

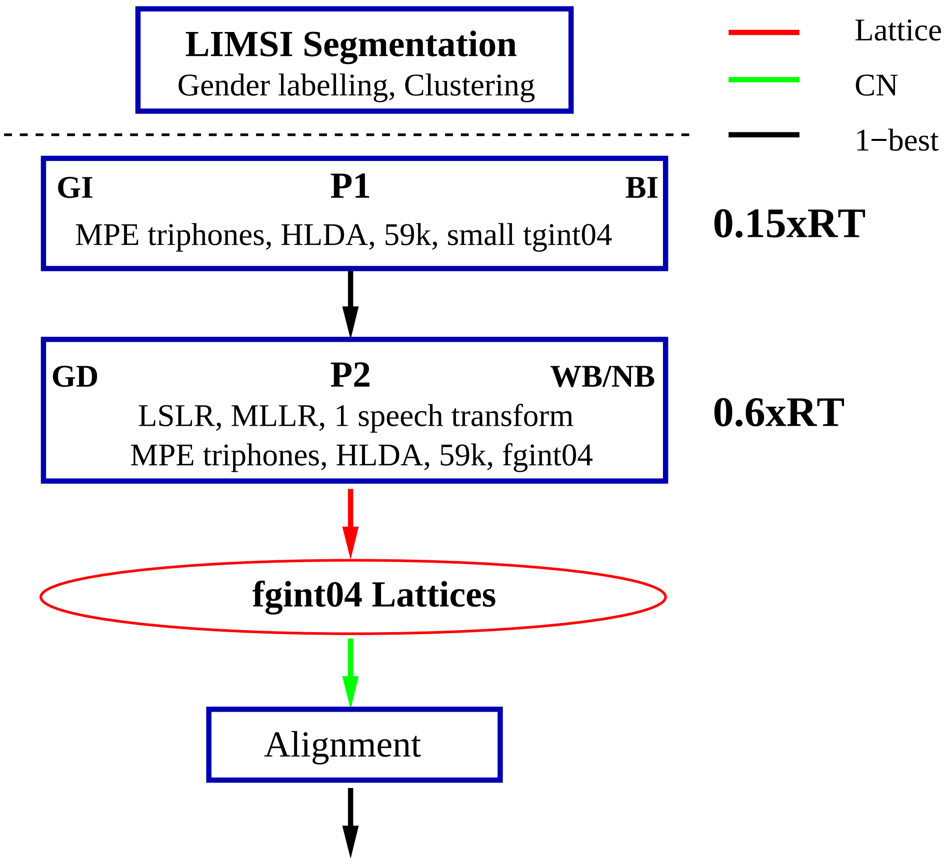
Performance of the Contrast system in comparison with the RT03 10×RT system.

- 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`

CU-HTK RT04 1xRT System Structure

- Fast version of P1+P2 from 10xRT
  - very fast P1 (0.15xRT)
  - P1 WER does not affect P2 WER much
  - used same P2 gender dependent acoustic models + adaptation
  - smaller LMs in P1/P2





RT03/04 CU BN-E Performance Comparison

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.

- consistent improvements from new models/system structure
- 10×RT: 26% relative error reduction on progress set
- 1×RT: 30% relative error reduction on progress set

SuperEARS : Cross-site System

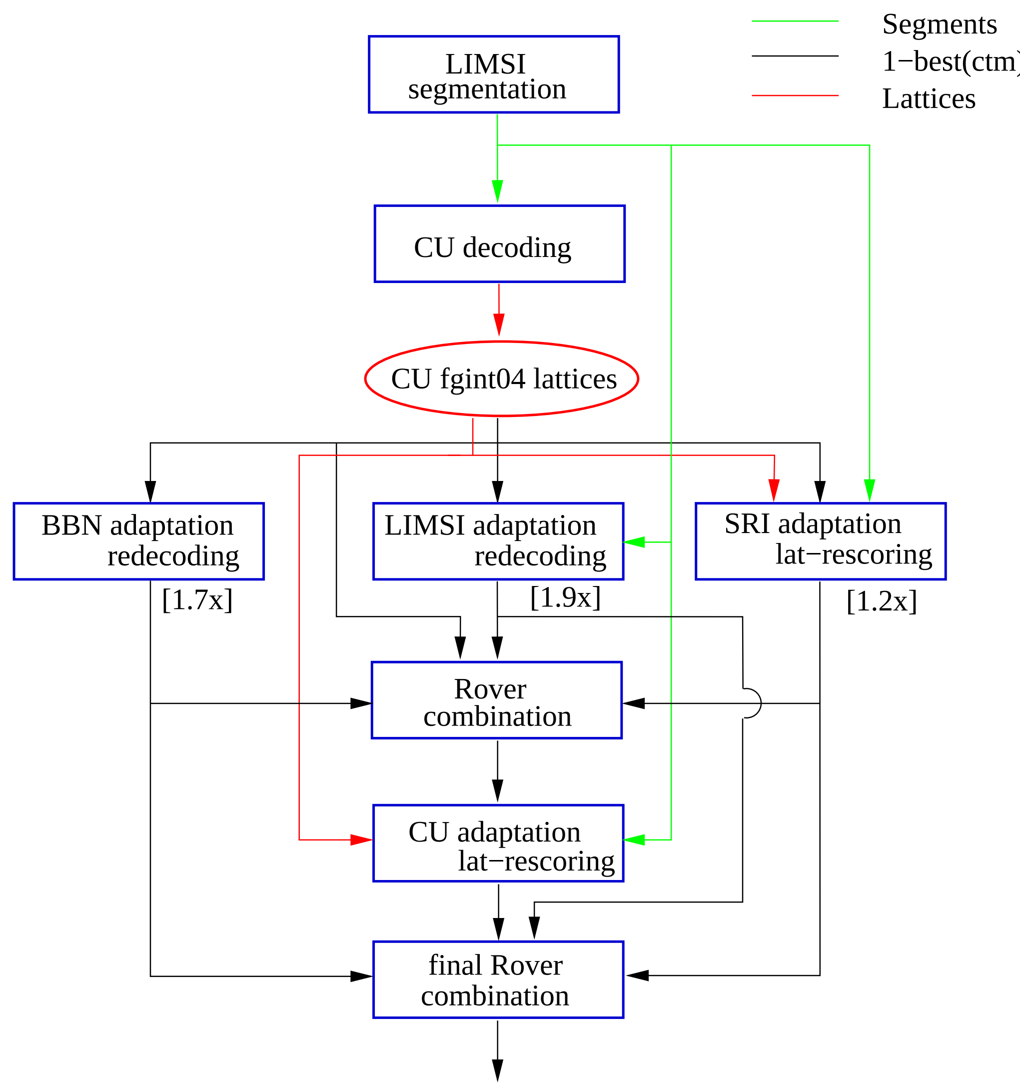
- Cross-site Combination
  - exploit & combine strengths of various EARS systems
  - implicit and explicit combination
  - need efficiency for $< 10\times$RT runtime constraint
  - robustness to test-set variability

- Initial feasibility test: 25% WER reduction from simple combination

System		Run-time	%WER on dev04
CU	May04	$< 10\times$RT	12.6
BBN	May04	$< 10\times$RT	12.7
LIMSI	RT03	$< 10\times$RT	13.7
SRI	May04	$< 10\times$RT	13.8
CU+BBN+LIMSI+SRI		$< 40\times$RT	**9.5**

SuperEARS System Structure

- LIMSI segmenter **[0.1xRT]**
- CU lattice generation **[3.5xRT]**
- 3-way rescoring/redecoding
  - BBN adapt/redecoding
  - LIMSI adapt/redecoding **[4.8xRT]**
  - SRI adapt/lat-rescoring
- ROVER combination
- CU final adaptation/rescoring **[0.9xRT]**
- Final ROVER combination



SuperEARS System Performance

Stage	%WER			
	eval03	dev04	dev04f	eval04
CU-lat	8.6	11.1	15.9	13.6
BBN-decode	8.1	9.8	14.3	12.8
LIMSI-decode	8.2	10.5	15.9	14.0
SRI-rescore	7.9	9.7	16.5	14.6
ROVER-superv	7.1	8.9	13.9	12.2
CU-adapt	7.6	9.6	14.3	12.8
ROVER-final	6.7	8.3	13.4	11.6

- Final output 1.9%-2.5% lower WER than lattice generation
- Performance of individual components varies across test-sets
- SuperEARS output very robust to component test-set variation

Performance Comparison

- SuperEARS system showed
  - 1.0% abs lower WER than single best system on `eval04`
  - 0.8% abs lower WER than best single system on progress set

- Compare with simple ROVER combination of three RT04 primary $< 10\times$RT systems
  - same performance as SuperEARS system at 3 times the run-time

System	Run-time	%WER		
		`dev04`	`dev04f`	`eval04`
BBN+LIMSI	$< 10\times$RT	9.4	14.0	12.7
CU (primary)	$< 10\times$RT	10.0	14.7	12.6
SRI	$< 10\times$RT	10.9	18.0	15.0
BBN+LIMSI+CU+SRI	$< 30\times$RT	8.2	13.5	11.6
SuperEARS	$< 10\times$RT	8.3	13.4	11.6

Conclusions

- For the RT04 BN 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
  - more language model training data/increased size
  - use of two segmentations
- Optimised 1×RT system including adaptation
  - 1×RT system WER 0.8-0.9% abs **lower** than RT03 10×RT system
- SuperEARS system
  - large gains possible by simple combination of multiple BN systems
  - efficient use of hybrid framework of lattice rescoring and re-decoding
  - 1% abs better than single best system on `eval04`