Design of Fast LVCSR systems

Gunnar Evermann, Phil Woodland &
Rest of the CU-HTK STT team

October 21st 2003

Cambridge University Engineering Department

Overview

- EARS project
- Background: LVCSR evaluation systems
- System structure for 10xRT
- Review of previous 10xRT CU-HTK systems
- Speed/accuracy trade-off
- 2003 system results
- Conclusions

DARPA EARS Programme

- Effective, Affordable, Reusable Speech-to-text (EARS) 5-year programme
- Work on two English corpora:
  - Broadcast News (BN, formerly known as Hub4) for first 3 years
    144h training from CNN, ABC, etc.; typical WER: 15%
  - Conversational Telephone Speech (CTS, aka Switchboard or Hub5)
    360h training, phone conversations on assigned topics; typical WER: 25%
- Aims are:
  - Very significant reduction in WER
  - Substantial speedup of current systems
  - Final targets: 5% WER at 1x Real Time
  - Generate metadata (speaker labels, punctuation, etc.)
    words + metadata = *Rich Transcription*
  - Port English work to Mandarin and Arabic

EARS at CUED: HTK-RAT

- CUED HTK Rich Audio Transcription project
- One of 3 funded Speech-to-Text (STT) teams: SRI, BBN+LIMSI, CUED
- 3 staff + 6 RAs + 6 PhD students + lots of computers
- Builds on previous work on BN (-1998) and CTS (1997-2002)
- Work on English BN and CTS and Mandarin CTS
- Participate in yearly Rich Transcription evaluations starting with RT02 (April 2002) and RT03 (March 2003)
- Benefits for rest of MIL: compute infrastructure, access to large amounts of data, complete state-of-the-art LVCSR systems

Building Fast Systems

- Recently increased interest in making state-of-the-art eval systems fast and thus feasible for practical use
- Several sites have had systems for 10xRT BN and unlimited CTS for some time (Primary condition for RT03)
- RT04/05 will be much more difficult with limits on CTS and $<$5xRT BN
- CTS is harder, due to higher task & system complexity
- Prepare for future evals and concentrate on appropriate techniques
- Build and submit prototype systems (10xRT CTS in RT02 & RT03)

Background: Typical LVCSR evaluation systems

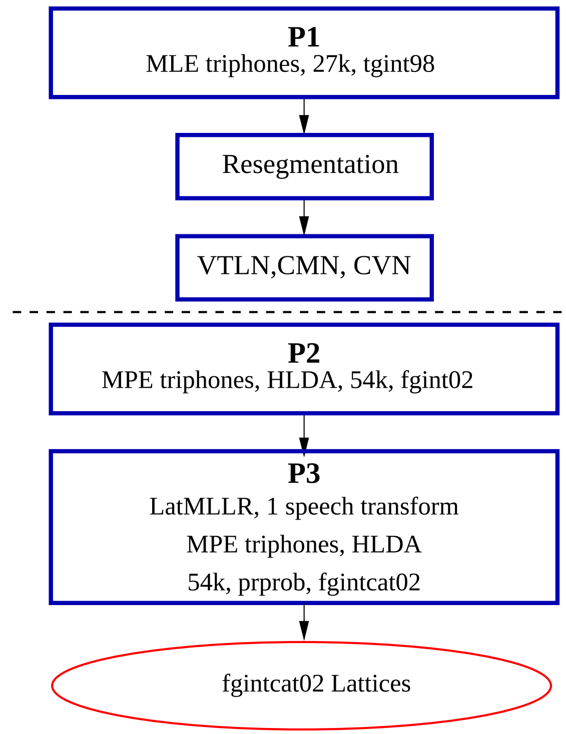
- Perform multiple decoding passes, to allow for complex models/adaptation
- Later passes often operate on lattices
- Combine output from multiple branches with different models
- Employ word N-gram + class N-gram language models
- Use PLP, VTLN, CMN, CVN, HLDA, MPE, SAT, SPron, latMLLR, FV, CNC...
- Very slow: 200-2000 xRT

Background: 2002 CU-HTK CTS system (320xRT)

- System has 2 stages: Lattice generation and multi-model lattice rescoring

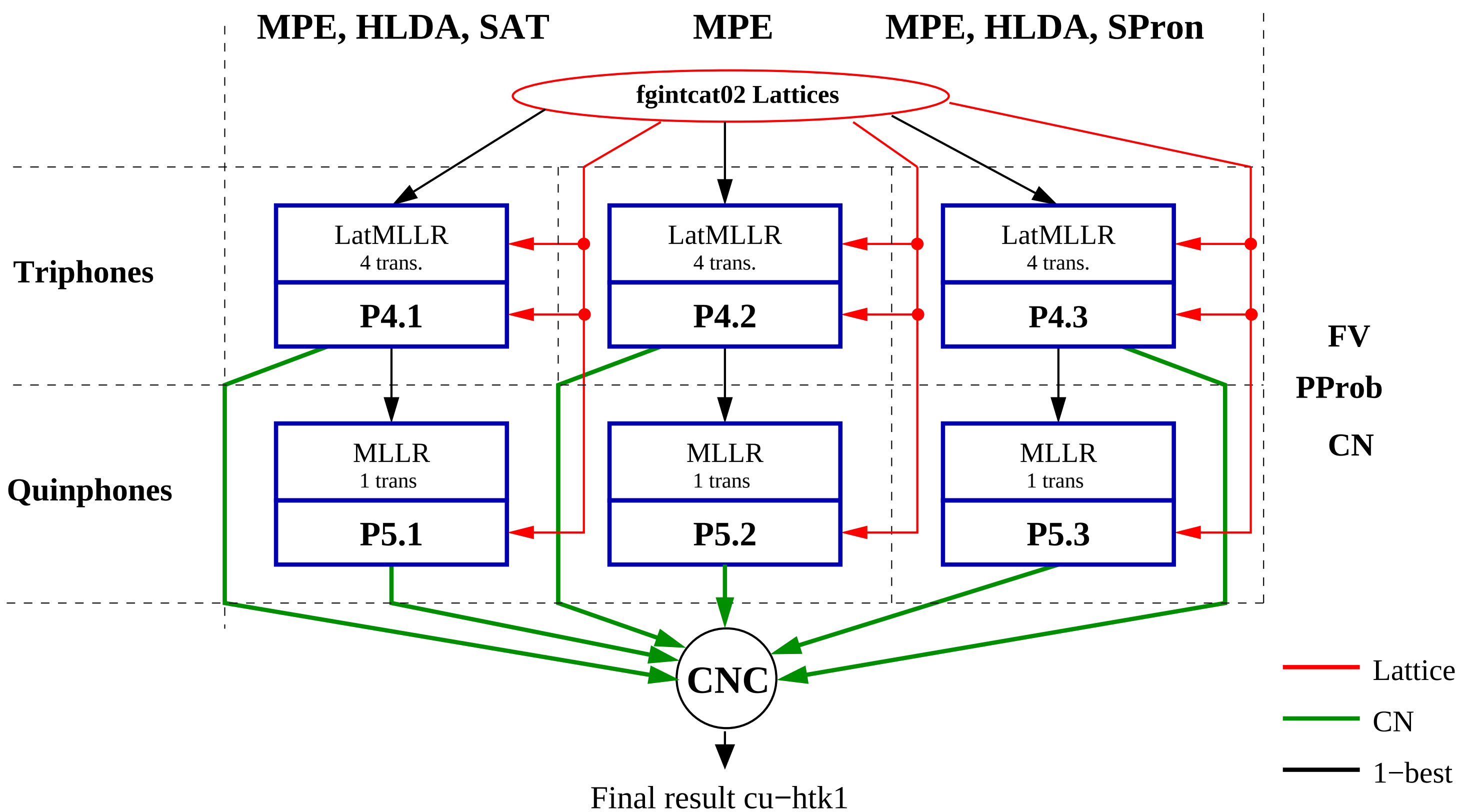
Lattice Generation

- Aim is to restrict search space for rescoring stage
- Use 3 full decoding passes
- P1 (initial transcription): only used to improve segmentation and as VTLN supervision
- P2 (supervision): produce supervision for MLLR
- P3 (lattice generation): generate very large lattices



Background: 2002 CU-HTK CTS system (320xRT)

- Lattice Rescoring in 3 branches: SAT, non-HLDA and SPron
- P4: triphones; P5: quinphones

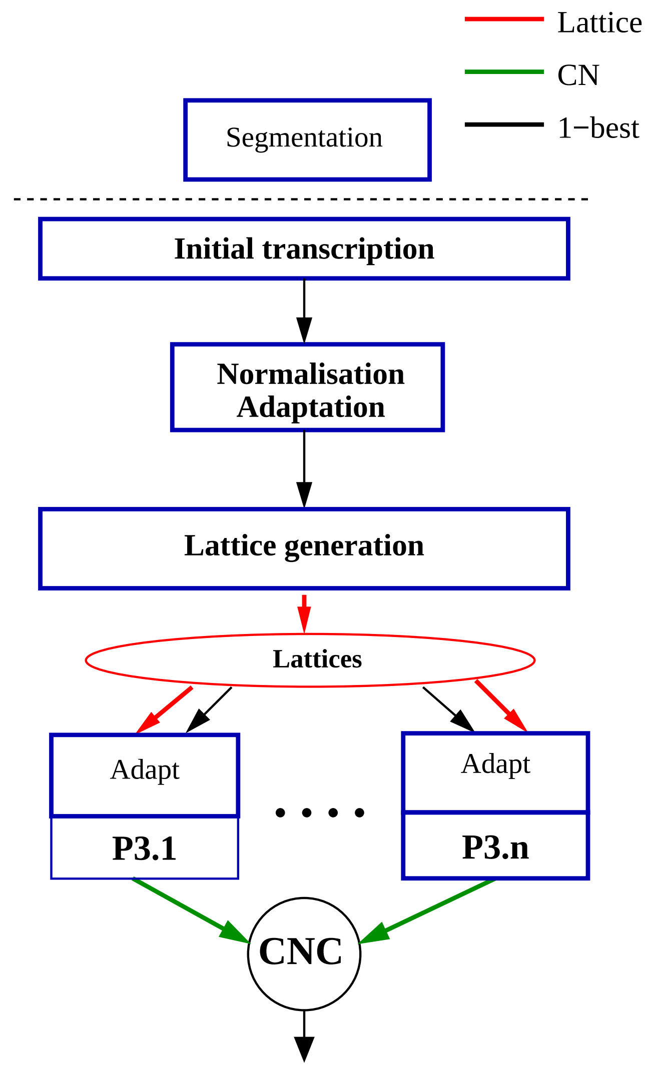


10xRT systems: Design decisions

- Use same system structure for BN and CTS
- No quinphones (only marginal gains at very high cost)
- Generate lattices in 2 passes, i.e. use P1 as supervision for VTLN and MLLR
- Simplify all adaptation processes (e.g. fewer transforms/passes)
- Tighter search parameters in all passes (large system is very conservative)
- Use only 2 rescoring branches, i.e. 2-way system combination

General system structure for 10xRT (BN/CTS)

- Segmentation
- Initial transcription **1xRT**
- Normalisation (re-segment, VTLN, etc.) Adaptation **0.5xRT**
- Lattice generation with word+class LM **4xRT**
- Lattice rescoring: for each model set: **2xRT**
  - Adaptation: MLLR (1-best + lattice), FV
  - Lattice rescoring
  - Confusion network generation
- System combination



Choosing Rescoring Model Sets

- Due to runtime constraints only 2-way system combination feasible
- Four MPE triphone sets were built for the CTS system:

A: SAT HLDA **B:** HLDA
C: SPron HLDA **D:** non-HLDA

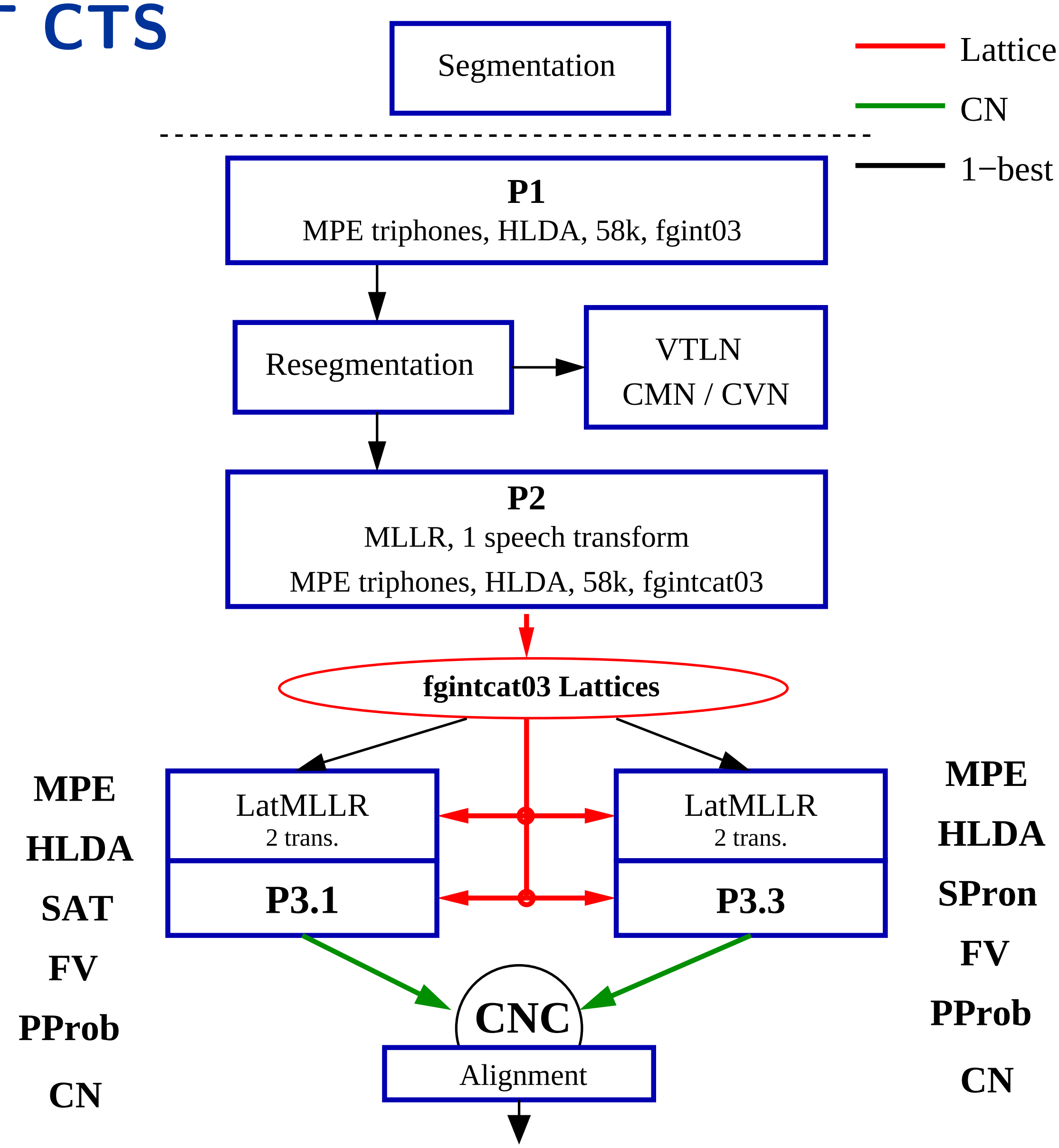
Results of pairwise system combination using CNC:

System	A	B	C	D
	23.0	23.6	23.4	24.8
+A		23.1	**22.6**	22.7
+B			22.9	23.3
+C				22.8

Individual Systems and pairwise combination
%WER on cts-eval02 after lattice-MLLR/FV and CN

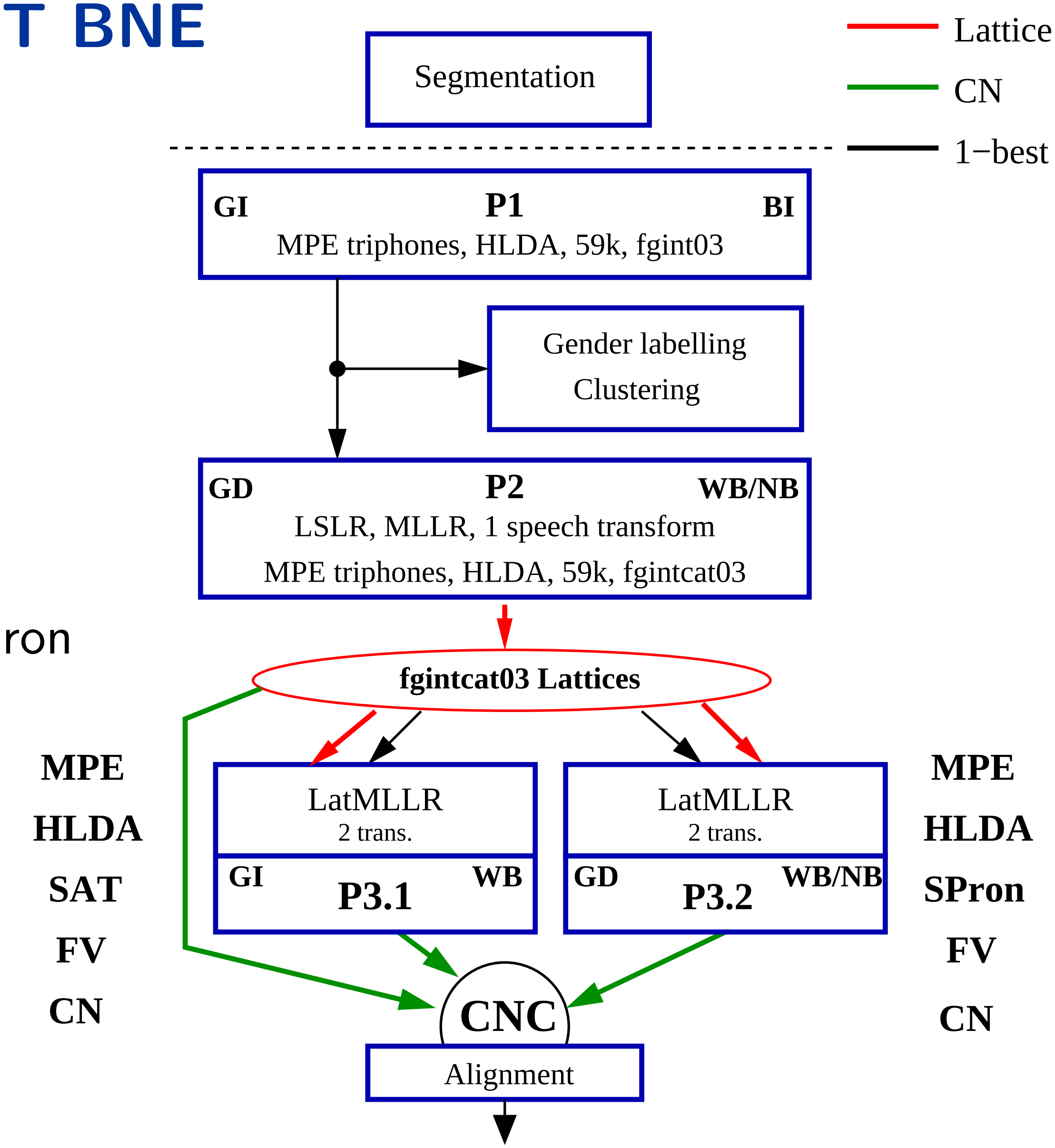
2003 System structure 10xRT CTS

- Automatic segmentation
- Use new models from full system
- All models use MPE, HLDA
- P2: use HLDA model for latgen
- Use lattice MLLR and full-variance
- Selected most effective 2-way combination (SAT & SPron)



2003 System structure 10xRT BNE

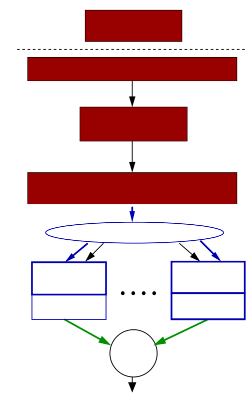
- Automatic segmentation
- Speaker clustering
- All models use MPE, HLDA
- P2:gender-/bandwidth-specific MPron
- P3:
  - SAT for wideband
  - SPron for M/F and NB/WB
- 3-way system combination



Previous work

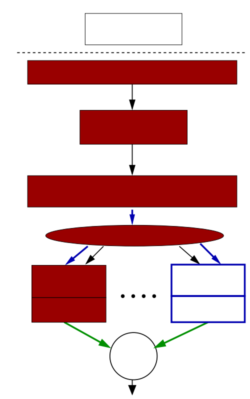
10xRT 1998 BN CUHTK-Entropic system:

- Single branch, two pass system, no lattice rescoring
- Automatic segmentation, speaker clustering
- Purpose-built acoustic models



10xRT 2002 CTS CUHTK system:

- Simple three pass system, based on full 320xRT system.
- Used models from full system (incl. 4 year old Pass 1 models!)
- No system combination



Software used

- All decoders use single tree static search network with multiple LM state-dependent tokens
- lattice generation (P1,P2):
  - Based on Entropic decoder, optimised for speed
  - Fast Gaussian computation
  - word-**pair** approximaton for fast trigram search
- lattice rescoring (P3.x):
  - More flexible (general adaptation, pronprobs, etc.)
  - Can rescore with arbitrary deterministic finite-state "LM" (e.g. lattices)
  - Generates lattice on output
- Rest is vanilla HTK3 plus new adaptation code

Optimising speed/accuracy trade-off

- Accuracy of intial pass has little influence on overall result

P1 speed	WER		
xRT	P1	P2 trigram	P2 fourgram
0.48	37.4	26.3	25.5
0.83	35.2	26.3	25.4
1.50	34.4	26.1	25.2

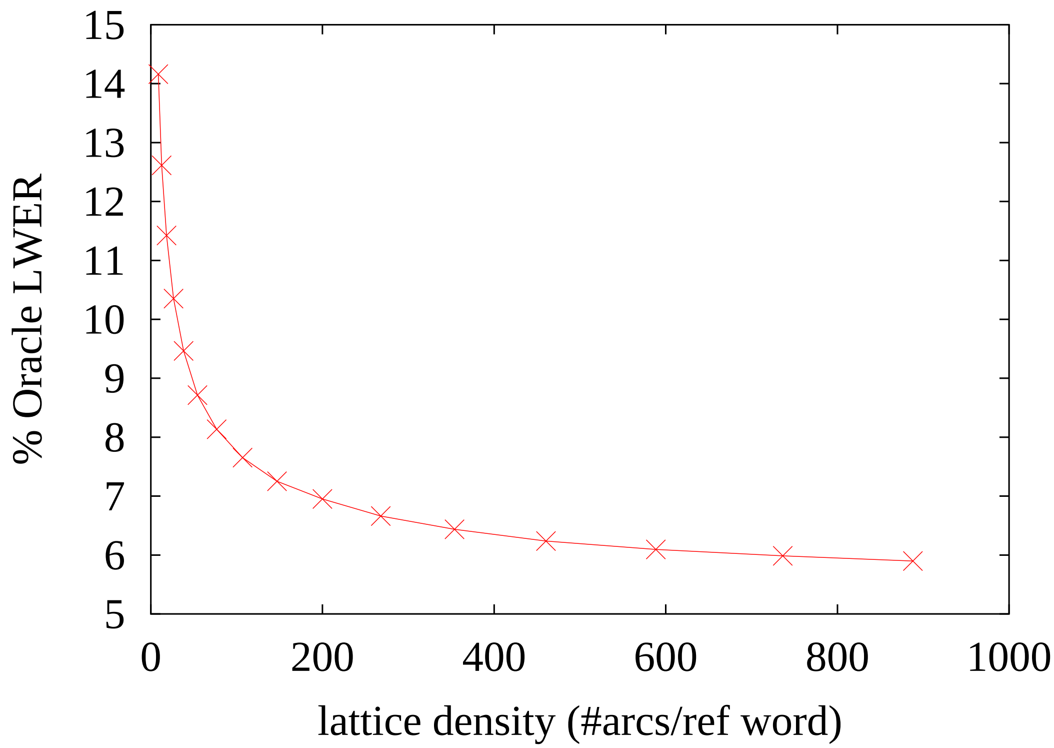
P1 speed-accuracy trade-off (CTS eval02)

- Choose middle operating point for safety (avoid failures)

Tuning lattice size

- Larger rescoring lattices are more likely to contain the correct answer...

use "Oracle" to find path with lowest WER (compared to reference) in lattice

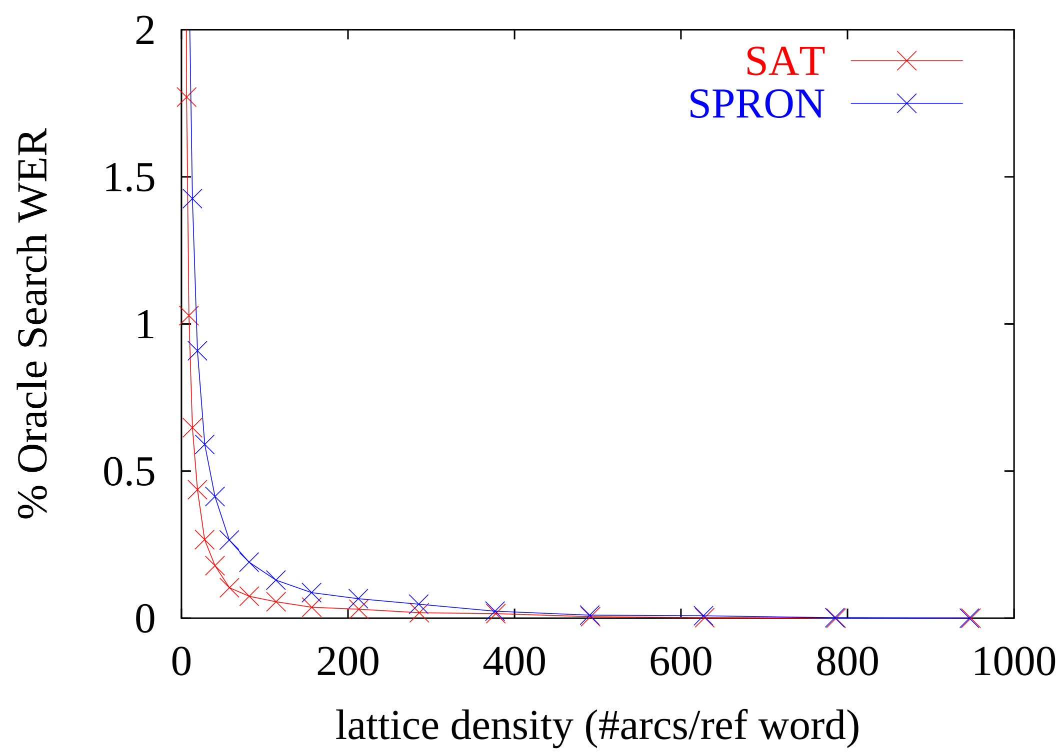


Oracle word error rate against lattice density (CTS eval02, P2-fg)

Tuning lattice size (cont'd)

- ...but we probably won't find it anyway:

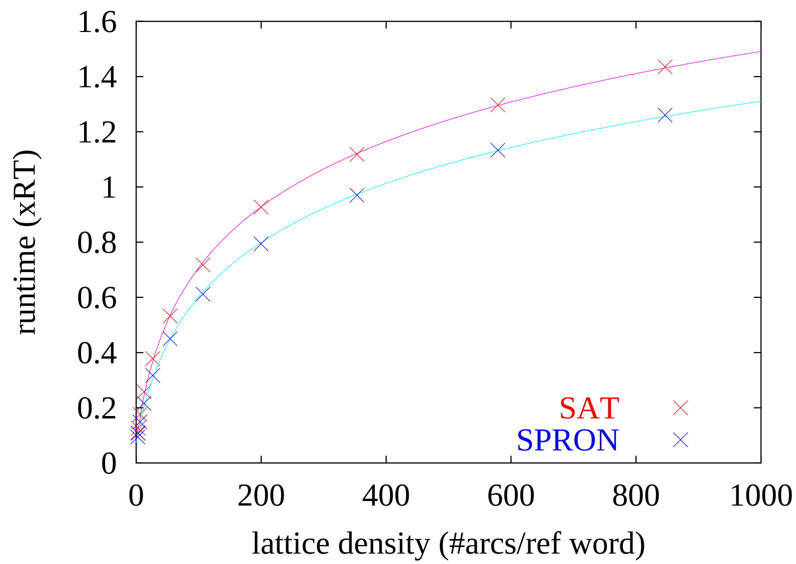
Oracle Search WER: rescore big lattices and take result as "reference" for oracle



Lattice search word error rate against lattice density (CTS eval02, P2-fg)

Predicting rescoring time

- To hit xRT target it is useful to predict rescoring time (P3) and prune lattices accordingly



Rescoring runtime against lattice density & fit of $a + b\log(x + c)$ (CTS eval02)

- Curves are roughly log-shaped

- Reason: size of search network grows logarithmically with lattice density!

How to make it run fast

- All decoding parameters were carefully chosen to stay in compute budget
- Important to limit worst-case behaviour (max model beams, lattice pruning)
- Simplify adaptation, e.g. use 2 speech transforms instead of 4
- Buy many fast computers! For eval and, more importantly, experiments. CUED compute infrastructure:
  - cluster of 40 IBM x335 dual Xeons
  - SunGrid batch queuing system (720k jobs since Nov'02)
  - for eval runs: keep all data local, use 20 fastest single CPUs (2.8GHz) turn around for 6 hour CTS set: 3 hours
- Avoid excessive overhead (e.g. reading LMs) by running on large subsets, e.g. complete BN shows or sets of several CTS sides

CTS: Final results on eval03

	Swbd	Fisher	Total
P1	39.0	29.7	34.5
P2	29.4	20.9	25.3
P3.1-cn	26.0	18.8	22.5
P3.3-cn	26.3	18.9	22.7
final	25.5	18.4	22.1

%WER on eval03 for 2003 10xRT system

- The system ran in 9.21 xRT (on the dev set: 9.17xRT)
- The confidence scores have an NCE of 0.318

CTS: Progress over last year

CUED internal aims were:

- Automate running of multi-pass 10xRT system
- Outperform last year's full 320xRT system in 10xRT
- Narrow gap between full and fast systems

	Swbd1	Swbd2	Cellular	Total	fast gap
320xRT 2002†	19.8	24.3	27.0	23.9	
10xRT 2002†	22.3	27.7	31.0	27.2	+14%
190xRT 2003	18.6	22.3	23.7	21.7	
10xRT 2003	19.9	23.5	25.8	23.3	+7%

%WER on eval02 for full and fast systems

†: using manual segmentation

gap on eval03 is 7%, on the progress set it is 5%.

BN: Final results on eval03

	WER
P1	**14.6**
P2.fgintcat	11.9
P2.fgintcat-cn	11.6
P3.1-cn†	11.4
P3.3-cn	11.4
final	**10.7**

%WER on eval03 for 2003 10xRT system

† wideband only, narrowband from P3.3

- P1 ran in 0.88 xRT – submitted as contrast, not an optimised 1xRT system!
- The full system ran in 9.10 xRT
- The confidence scores have an NCE of 0.412

BN: System combination

- Combination in BN system is more complicated than CTS, as we had no BN narrow-band SAT models
- Employ 3-way combination (P2, SAT, SPron) for wideband, 2-way (P2, SPron) otherwise.
- Mismatch of posterior distributions due to lattice sizes (P2 are much bigger than P3)
- Ongoing work: Investigate mapped posteriors, system weights etc.

Conclusions

- BN: rebuilt setup and constructed state-of-the-art 10xRT system
- CTS: good improvements over RT02 systems
- Narrowed gap between 100+ xRT and 10xRT considerably
- Infrastructure for quick-turnaround *system* tests (vs. single *model* experiments)

Future Work

- Optimise models (HMMs and LMs) for fast systems
- Fast versions of VTLN and MLLR
- Adaptive optimisation of system structure