CU-HTK Fast System Description

Gunnar Evermann, Do Yeong Kim, Lan Wang, Phil Woodland
+ Rest of the HTK STT team

May 19th 2003

Cambridge University Engineering Department

Overview

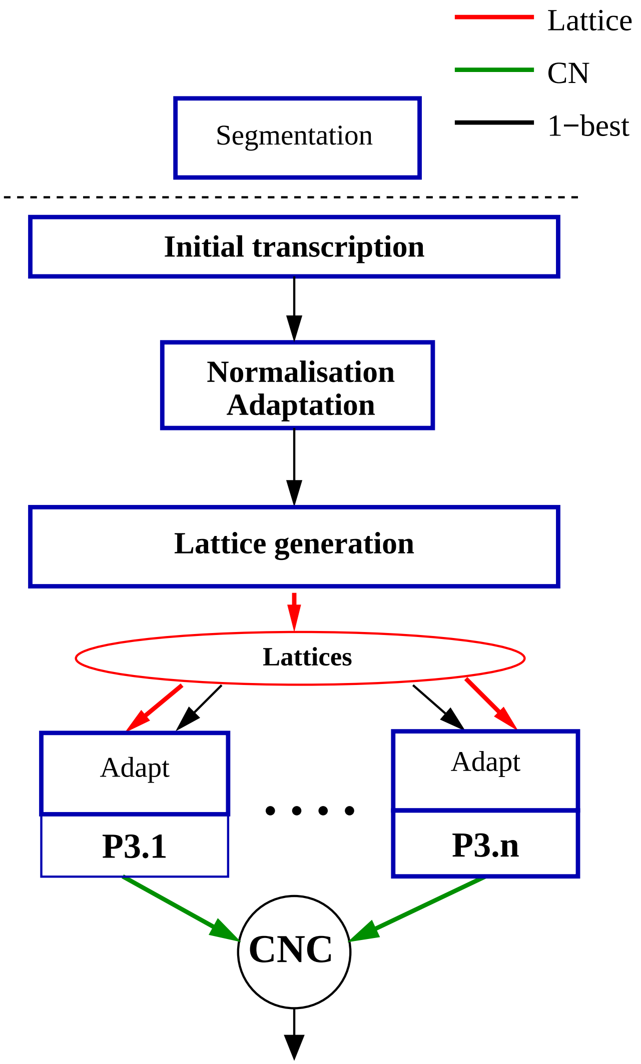
- Introduction
- System structure for 10xRT
- Review of previous 10xRT CU-HTK systems
- 10xRT system development
- 2003 system results
- Conclusions

Introduction

- 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 RT02)
- 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)

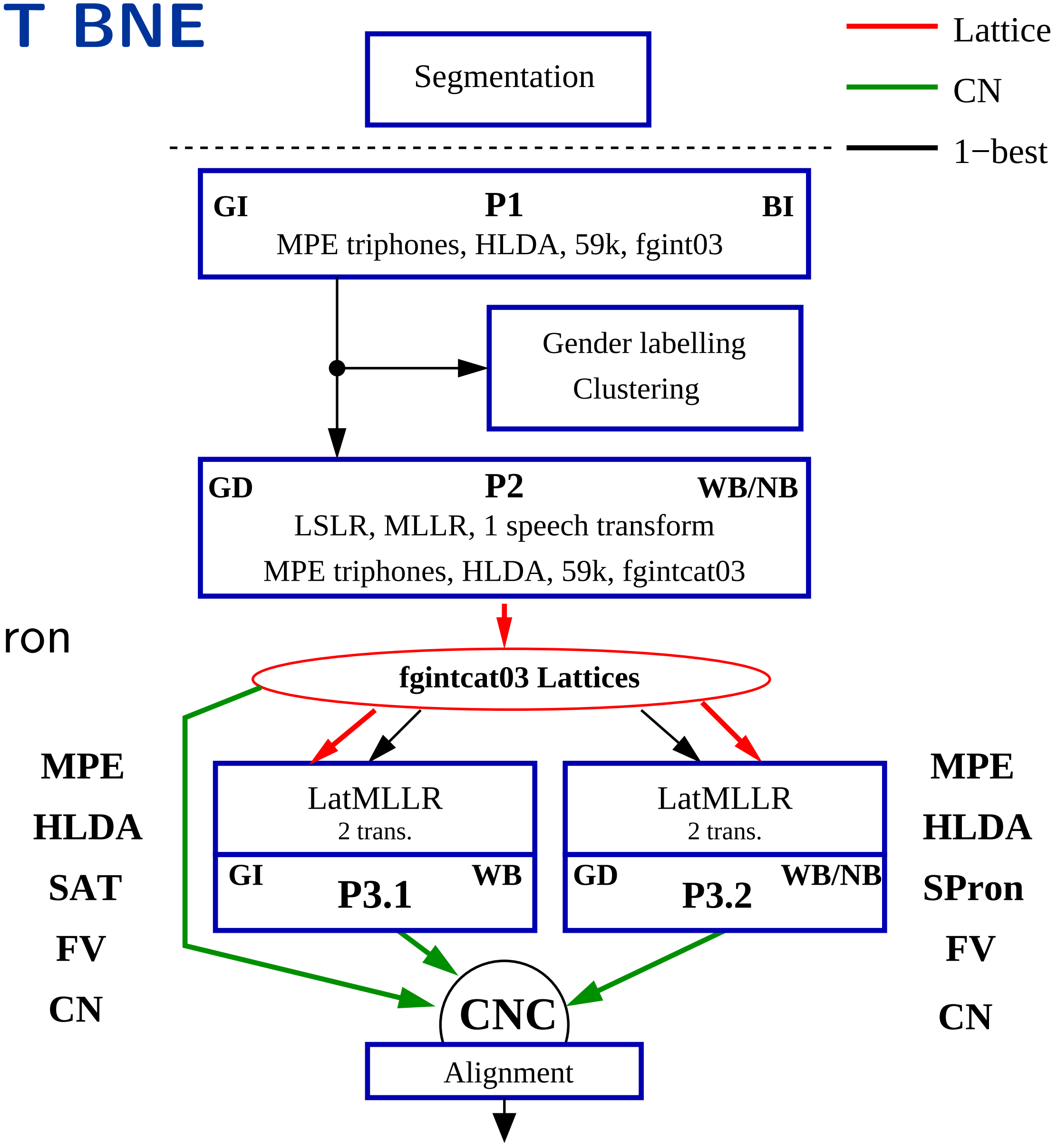
General system structure for 10xRT (BN/CTS)

- 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



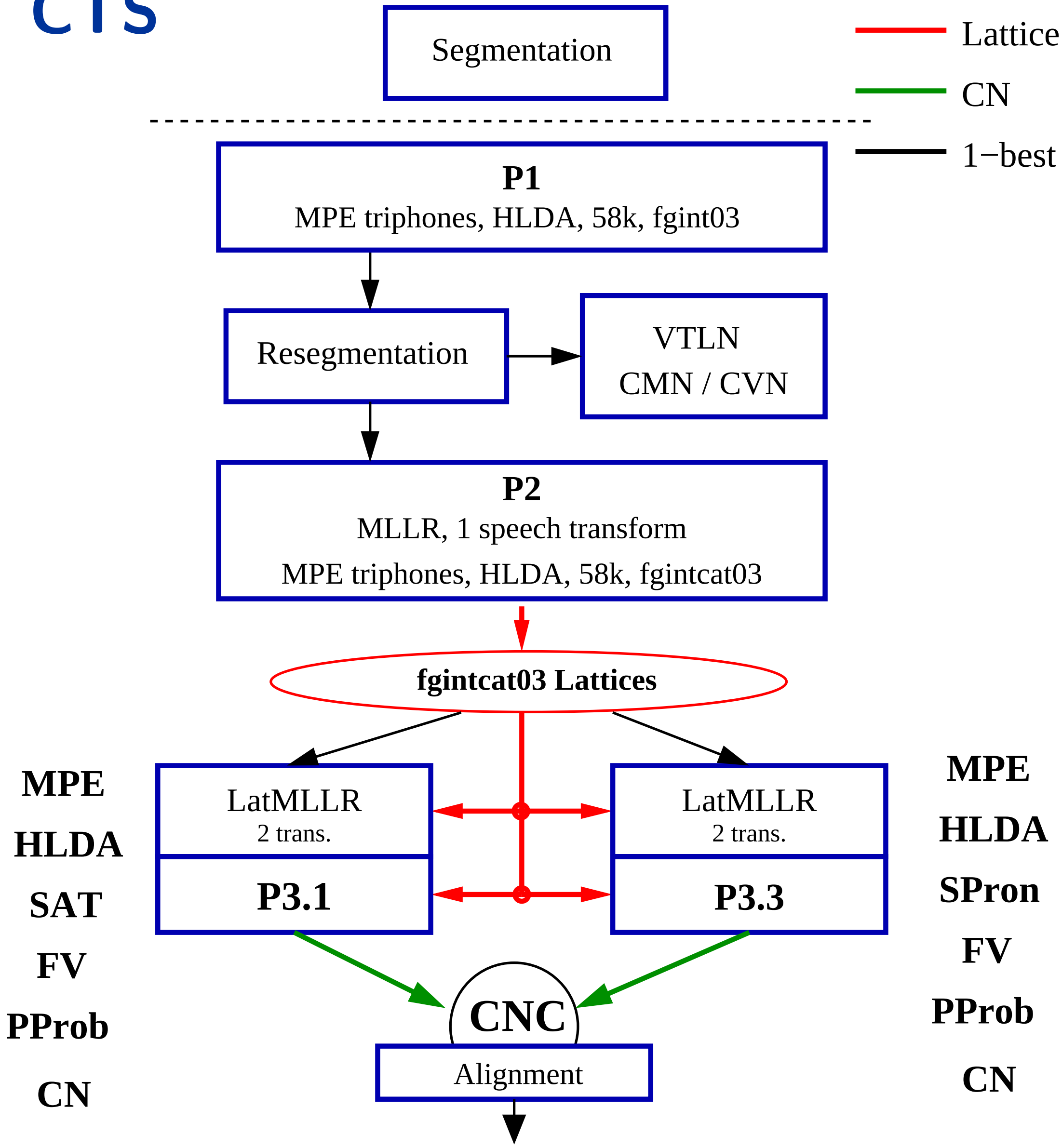
2003 System structure 10xRT BNE

- 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



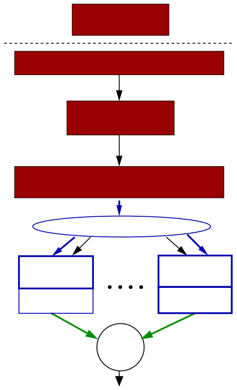
2003 System structure 10xRT CTS

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



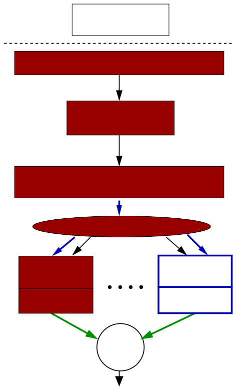
Previous work

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, built in a few of days based on full 320xRT system.
- Used models from full system (incl. 4 year old Pass 1 models!)
- No system combination

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 IBM x335 dual Xeons
  - SunGrid batch queuing system (400k 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: Development results on eval02

	Swbd1	Swbd2	Cellular	Total
P1	28.7	36.3	40.2	35.5
P2	22.4	26.8	29.8	26.6
P3.1-cn	20.4	24.0	26.1	23.7
P3.3-cn	20.4	24.3	26.6	24.0
final	19.9	23.5	25.8	23.3

%WER on eval02 (automatic segmentation) for 2003 10xRT system

- The system ran in 9.17 xRT
- The confidence scores have an NCE of 0.295

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
- The confidence scores have an NCE of 0.318

CTS: Progress over last year

CUED internal aims were:

- Automate running of 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: Development results on bndev03

	WER
P1	15.9
P2.fgintcat	13.1
P2.fgintcat-cn	12.8
P3.1-cn†	12.0
P3.3-cn	12.1
final	11.6

%WER on bndev03 for 2003 10xRT system

† wideband only, narrowband from P3.3

- The confidence scores have an NCE of 0.393

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 – submited 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 decoding parameters & structure