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

<table>
<thead>
<tr>
<th></th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>28.7</td>
<td>36.3</td>
<td>40.2</td>
<td>35.5</td>
</tr>
<tr>
<td>P2</td>
<td>22.4</td>
<td>26.8</td>
<td>29.8</td>
<td>26.6</td>
</tr>
<tr>
<td>P3.1-cn</td>
<td>20.4</td>
<td>24.0</td>
<td>26.1</td>
<td>23.7</td>
</tr>
<tr>
<td>P3.3-cn</td>
<td>20.4</td>
<td>24.3</td>
<td>26.6</td>
<td>24.0</td>
</tr>
<tr>
<td>final</td>
<td>19.9</td>
<td>23.5</td>
<td>25.8</td>
<td>23.3</td>
</tr>
</tbody>
</table>

%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

<table>
<thead>
<tr>
<th></th>
<th>Swbd</th>
<th>Fisher</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>39.0</td>
<td>29.7</td>
<td>34.5</td>
</tr>
<tr>
<td>P2</td>
<td>29.4</td>
<td>20.9</td>
<td>25.3</td>
</tr>
<tr>
<td>P3.1-cn</td>
<td>26.0</td>
<td>18.8</td>
<td>22.5</td>
</tr>
<tr>
<td>P3.3-cn</td>
<td>26.3</td>
<td>18.9</td>
<td>22.7</td>
</tr>
<tr>
<td>final</td>
<td>25.5</td>
<td>18.4</td>
<td>22.1</td>
</tr>
</tbody>
</table>

%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

<table>
<thead>
<tr>
<th></th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
<th>fast gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>320xRT 2002†</td>
<td>19.8</td>
<td>24.3</td>
<td>27.0</td>
<td>23.9</td>
<td></td>
</tr>
<tr>
<td>10xRT 2002†</td>
<td>22.3</td>
<td>27.7</td>
<td>31.0</td>
<td>27.2</td>
<td>+14%</td>
</tr>
<tr>
<td>190xRT 2003</td>
<td>18.6</td>
<td>22.3</td>
<td>23.7</td>
<td>21.7</td>
<td></td>
</tr>
<tr>
<td>10xRT 2003</td>
<td>19.9</td>
<td>23.5</td>
<td>25.8</td>
<td>23.3</td>
<td>+7%</td>
</tr>
</tbody>
</table>

%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

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>15.9</td>
</tr>
<tr>
<td>P2.fgintcat</td>
<td>13.1</td>
</tr>
<tr>
<td>P2.fgintcat-cn</td>
<td>12.8</td>
</tr>
<tr>
<td>P3.1-cn†</td>
<td>12.0</td>
</tr>
<tr>
<td>P3.3-cn</td>
<td>12.1</td>
</tr>
<tr>
<td>final</td>
<td>11.6</td>
</tr>
</tbody>
</table>

%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

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>14.6</td>
</tr>
<tr>
<td>P2.fgintcat</td>
<td>11.9</td>
</tr>
<tr>
<td>P2.fgintcat-cn</td>
<td>11.6</td>
</tr>
<tr>
<td>P3.1-cn†</td>
<td>11.4</td>
</tr>
<tr>
<td>P3.3-cn</td>
<td>11.4</td>
</tr>
<tr>
<td>final</td>
<td>10.7</td>
</tr>
</tbody>
</table>

%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 decoding parameters & structure