Woodland et al.: English CTS Systems

2003 system performance

- Revisited language models
- SPFon experiments
- Additional acoustic training data
- SAT experiments
- Lattice generation for MPE training
- Variable number of Gaussians
- Revisited transcriptions
- Revisited language models
- Automatic segmentation

New/Revised components

Training and test data sets
2002 unlimited computation system

English CTS Development

Woodland et al.: English CTS Systems
Woodland et al.: English CTS Systems

2002 System - Lattice Generation

- HTK decoder HDecode
- Pronunciation probabilities
- Adaptation for P3 via lattice MLIR
- HLDAs
- 28 mixture components (28 mix)
- MPE triphone models for P2/P3
- MLE P1 models

Use confusion networks to represent each rescoring pass output & confusion network combination for highest posterior probability words and confidence scores

Rescore the lattices with adapted triphone and quinphone models

Expands the lattices to 4-gram plus bigram category model

Generates lattices with adapted triphone models and a bigram LM

Initial passes generate transcriptions for VTLN & initial adaptation

(PLP, side-based CMN/CVN + 1st/2nd + 3rd & HLDA to 39 dims)

Assumes manual segmentation into turns
Results on eval02 set

<table>
<thead>
<tr>
<th></th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 trans for VTLN</td>
<td>35.6</td>
<td>44.6</td>
<td>30.9</td>
<td>35.6</td>
</tr>
<tr>
<td>P2 trans for MLLR</td>
<td>24.6</td>
<td>28.3</td>
<td>30.8</td>
<td>26.0</td>
</tr>
<tr>
<td>P3 lat gen</td>
<td>22.5</td>
<td>23.2</td>
<td>26.3</td>
<td>23.7</td>
</tr>
<tr>
<td>P4.1 SAT tri</td>
<td>21.6</td>
<td>25.5</td>
<td>28.6</td>
<td>24.3</td>
</tr>
<tr>
<td>P4.2 non-HLDA tri</td>
<td>22.3</td>
<td>27.4</td>
<td>31.3</td>
<td>26.9</td>
</tr>
<tr>
<td>P4.3 SPron tri</td>
<td>21.5</td>
<td>26.6</td>
<td>31.2</td>
<td>26.2</td>
</tr>
<tr>
<td>P5.1 SAT tri</td>
<td>21.5</td>
<td>26.4</td>
<td>30.9</td>
<td>25.9</td>
</tr>
<tr>
<td>P5.2 non-HLDA tri</td>
<td>22.9</td>
<td>28.0</td>
<td>31.0</td>
<td>27.2</td>
</tr>
<tr>
<td>P5.3 SPron tri</td>
<td>22.5</td>
<td>27.4</td>
<td>30.8</td>
<td>26.7</td>
</tr>
<tr>
<td>P5.4 SAT tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
<tr>
<td>P5.5 SAT tri</td>
<td>21.5</td>
<td>26.4</td>
<td>28.8</td>
<td>25.3</td>
</tr>
<tr>
<td>P5.6 Sat tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
<tr>
<td>P5.7 SAT tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
<tr>
<td>P5.8 SAT tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
<tr>
<td>P5.9 SAT tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
<tr>
<td>P5.10 SAT tri</td>
<td>21.5</td>
<td>26.6</td>
<td>29.1</td>
<td>25.4</td>
</tr>
</tbody>
</table>

%WER on eval02 set of 2002 system, manual segmentation

Final confidence scores have NCE 0.289

2002 system – Rescoring & Combination

Woodland et al.: English CTS Systems

Rich Transcription Workshop 2003

Cambridge University Engineering Department
Training and Test Data Sets

Woodland et al.: English CTS Systems

Automatic Segmentation

Development test sets

Phase 2 data released by BBN (CTRANS transcribed)

HTKTRAIN3B 360 hr set. As above plus extra Switchboard Cellular I and Swbd2

HTKTRAIN3A 240 hr set. As above plus extra Switchboard I from final

LDC cell I corpus (without dev01/eval01 sides) extra 17 hrs of data

HTKTRAIN2 244 hrs Switchboard (Swbd1), 17 hrs CallHome English (CHE) +
A mistake in the Switchboard training transcriptions used in building all CUHTK CTS systems since 2000 was discovered.

Error in processing MSU Swbd training transcripts

Some fairly common words systematically deleted (3% of tokens)

Rebuilding acoustic models only, for 2002 10xRT system on eval02 (manual segs), Reduced WER by only 0.5% abs (27.2 to 26.7).

Rebuilding acoustic models only for 2002 10xRT system on eval02 (manual segs), Reduced WER by only 0.5% abs (27.2 to 26.7).

 Added 294 new conversation sides

Rebuilt transcriptions based on final version of MSU transcriptions

AFFECTED both acoustic models and LMs

Some fairly common words systematically deleted (3% of tokens)

Error in processing MSU Swedish training transcriptions

CTS systems since 2000 was discovered.
SAT/Adaptation Experiments

SAT tries to remove inter-speaker variability in training set by means of linear transform

Use constrained MLLR to generate a single transform per training side (can operate in feature space)

Interleave update of adaptation matrices and MLE HMM updates

Perform MPE training based on SAT models with fixed transforms

Integrate update of adaptation matrices and MLE HMM updates

Operate in feature space (can use constrained MLLR to generate a single transform per training side)

SAT tries to remove inter-speaker variability in training set by means of linear transform

SAT/Adaptation Experiments
In 2002 no re-alignment/regeneration of lattices during discriminative training

In 2001 re-generated model-marked lattices part way through MMI training

Now use heavily pruned training data bigram for word lattice generation

- In 2001 re-generated model-marked lattices part way through MMI training
- In 2002 no re-alignment/regeneration of lattices during discriminative training

This lattice regen/comb gives a further 0.5% abs improvement in WER

Larger lattices with pruned bigram reduce WER by about 0.3% abs

HLDA ML models to generate lattices reduces WER by about 0.4% abs

<table>
<thead>
<tr>
<th></th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-HLDA lattices</td>
<td>20.5</td>
<td>35.3</td>
<td>34.7</td>
<td>30.1</td>
</tr>
<tr>
<td>HLDA full bg + ug</td>
<td>20.4</td>
<td>34.7</td>
<td>34.3</td>
<td>29.7</td>
</tr>
<tr>
<td>HLDA pruned bg</td>
<td>20.0</td>
<td>34.4</td>
<td>34.0</td>
<td>29.4</td>
</tr>
<tr>
<td>MPE lattice regen/comb</td>
<td>19.4</td>
<td>34.0</td>
<td>33.6</td>
<td>28.9</td>
</tr>
</tbody>
</table>

MPE Training with Modified Lattices: Results

After 4 iterations of MPE training regenerate word and model-marked lattices

Use HLDA ML models to generate lattices (rather than non-HLDA lattices)

MPE lattice regen/comb

Use heavily pruned bigram scores in MPE training also
- better representation of comparable data
- larger "denominator" lattices
- In 2001 re-generated model-marked lattices part way through MMI training
In 2002 no re-alignment/regeneration of lattices during discriminative training

Now use heavily pruned training data bigram for word lattice generation

After 4 iterations of MPE training regenerate word and model-marked lattices

- Use HLDA ML models to generate lattices (rather than non-HLDA lattices)
Woodland et al.: English CTS Systems

Additional Acoustic Training

New Swbd2 data transcriptions provided by BBN (70 hours)

About 1% abs reduction in WER for MLE HMMs and 1.3% for MPE

Largest improvement for cellular data (2.2% abs) and Swbd2 data (1.4% abs)

Training of statistical model on decisions above

If words were observed in the training data

Systematically remove all pronunciation variants

SPron Dictionary

Woodland et al.: English CTS Systems

Modiﬁed procedure from 2002 CHTK CTS eval system (Hain, 2002)

Cambridge University Engineering Department Rich Transcription Workshop 2003 16
Woodland et al.: English CTS Systems

SPron experiments

Rebuilt SPron models with MPE lattice comb from MPron word lattices

Lattice combination helps 0.8% with SPron models built like this

Final MPron and SPron WERs very similar (SPron 1% abs better for MLE)

Lattice combination helps 0.8% with SPron models built like this

MPron

SPron

2003 Language models

Training data in 5 portions:

- Google data from U of Washington [6250]
- Swiss Transcriptions from BBN/CTRANS [0.9M]
- Cell transcripts [0.2M]
- Revised MSU Transcripts + CHE [3M]
- Broadcast news setup (BN transcripts from PSM; CNN data; TDT data)

% WER dev01 manual vs 2003 Trigram LM: 390 phrases, unadapted HLDAs

19.0 32.6 31.4 27.6 19.9 33.1 32.2 28.3
19.4 33.2 31.7 28.0 19.9 33.1 32.2 28.3
23.6 38.1 36.3 32.7
22.6 36.9 35.0 32.5
21.5 35.8 34.5 31.9
19.5 32.9 31.1 27.5

SPron experiments
Woodland et al.: English CTS Systems

2003 merged fgintcat has 4.3% rel reduction in PP over 2002 model. With cat models the difference is 3.5%

Effect of component 4-gram word LMs

<table>
<thead>
<tr>
<th>LMs</th>
<th>68.9</th>
<th>69.6</th>
<th>69.4</th>
<th>69.2</th>
<th>69.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>all minus BN+TD+CNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all minus che+swbd1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all minus swbd1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all minus cell1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all minus google</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all minus BN+TD+CNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2003 merged fgintcat has 4.3% rel reduction in PP over 2002 model. With cat source using Good-Turing (HTK HLM), "small text source trained using modified Kneser-Ney (SR1 LM); large text model"

Woodland et al.: English CTS Systems
2003 System Performance (Eval02)

<table>
<thead>
<tr>
<th></th>
<th>Swbd1</th>
<th>Swbd2</th>
<th>Cellular</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)-best CN</td>
<td>27.2</td>
<td>34.8</td>
<td>39.5</td>
<td>34.2</td>
</tr>
<tr>
<td>4 trans P3 lat.</td>
<td>23.6</td>
<td>28.9</td>
<td>31.7</td>
<td>28.4</td>
</tr>
<tr>
<td>P4.1 SAT tri.</td>
<td>21.1</td>
<td>25.1</td>
<td>27.6</td>
<td>24.8</td>
</tr>
<tr>
<td>P4.2 non-HLDA tri.</td>
<td>23.3</td>
<td>25.7</td>
<td>27.6</td>
<td>24.8</td>
</tr>
<tr>
<td>P5.1 SAT tri.</td>
<td>19.9</td>
<td>23.6</td>
<td>25.0</td>
<td>23.0</td>
</tr>
<tr>
<td>P5.2 non-HLDA quin.</td>
<td>21.2</td>
<td>25.1</td>
<td>27.6</td>
<td>24.8</td>
</tr>
<tr>
<td>P5.3 SPron quin.</td>
<td>20.1</td>
<td>23.6</td>
<td>25.0</td>
<td>23.0</td>
</tr>
<tr>
<td>Final NCE is 0.304</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[1\]-best CN: 1-gram.

\(1\)-best CN: 1-gram.

2003 System Part II
A number of changes and improvements have been made to the system although basic structure the same as 2002 system

- Automatic segmentation now gives only 0.6% increase in WER
- Revised Swbd transcriptions: 0.5% abs
- Manual segments with automatic segments. Approx 12% reduction in WER if use consistent
- On eval02 data got 23.9% WER in 2002 with manual segments: now 21.7%
- Approx 12% reduction in WER

**Conclusions**
revised language models: (guess) 0.2% abs on eval02 but expect more on Fisher data?

Overall the system ran in 187xRT

For Current Test Set error rate is 20.7%

For Progress Test Set (all Fisher) error rate is 17.4%

Many interesting things didn't make eval system this year

... –
– more advanced covariance modelling (e.g. extended MLLT)
– discriminative estimation of SAT transforms
– MMI/ME training of HLDA transforms

Fisher data?

Revised language models: (guess) 0.2% abs on eval02 but expect more on