New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

T. Hain, P.C. Woodland, G. Evermann, D. Povey

May 10th 2001

Cambridge University Engineering Department

ICASSP’2001 Salt Lake City
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

Outline

• Task and baseline system

• New features
  – Soft tying
  – MMIE based training
  – Pronunciation probabilities
  – Full variance transforms
  – Confusion networks

• Overall system: Structure and results

• Conclusions

• The CU-HTK 2001 system
Task Description

- System for NIST Hub5 evaluation

- Hub5 data
  - Switchboard
    Large corpus (>250 hours of speech)
  - CallHome English
    Small (<20 hours)
  - Presegmented at turn boundaries

- Hub5 Task
  - Transcription
  - Confidence scores

- Word error rates 20–40% and strongly test set dependent

- Test sets used here: eval98 and eval00
Basic System Features

- **Front-end**
  - Reduced bandwidth 125–3800 Hz
  - 12 modified PLP cepstral parameters + $c_0$ and 1st/2nd derivatives
  - Side-based cepstral mean and variance normalisation
  - Vocal tract length normalisation in training and test

- **Decision tree clustered, context dependent phone models:**
  triphone & quinphone HMMs

- **Multiple pronunciation dictionary** (based on LIMSI’93 + TTS)

- **N-gram backoff** (bigram, trigram & 4-gram) and class-category trigram LMs
  Interpolated Hub5 LMs and Broadcast News LMs

- **Speaker/channel adaptation:** MLLR
New Features and Changes

- New features
  - MMIE triphones and quinphones
  - Soft-tying of states
  - Unigram pronunciation probabilities
  - Full variance transform for speaker adaptation
  - Confusion networks

- Minor changes to the overall system
  - Increased vocabulary size (27k → 54k)
  - Updated transcriptions for acoustic and LM training
  - Increased amount of acoustic training data
  - Data weighting
Baseline Model Training

- Full training set (h5train00)
  - 248 hours Swbd-I
  - 17 hours CallHome English (CHE)
  - Manually corrected MSU transcriptions from Jan 2000
  - Removed silence at segment boundaries

- MLE triphone models
  - Initial models trained on 68 hour subset using VTLN data
    (6k states / 12 mix)
  - Extended training using 265 hour set h5train00 (16 mix)

- MLE quinphone models
  - ±2 phone context + word boundary clustering on h5train00 VTLN data
  - Trained up to 16 mix (9k states)
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

**Soft-tying**

- Sharing Gaussians between a fixed set of states (Luo et al. 1998)
  - State-specific mixture weights

- Simplified approach
  - Single Gaussian versions of full HMMs
  - Use Gaussian overlap distance to find the nearest 2 states
  - Add Gaussians from nearest states with uniform weighting and retrain

<table>
<thead>
<tr>
<th>System Type</th>
<th>Triphones</th>
<th>Quinphones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd-II</td>
<td>CHE</td>
</tr>
<tr>
<td>GI</td>
<td>42.5</td>
<td>47.7</td>
</tr>
<tr>
<td>ST/GI</td>
<td>42.1</td>
<td>47.4</td>
</tr>
<tr>
<td>ST/GD</td>
<td>41.4</td>
<td>47.0</td>
</tr>
</tbody>
</table>

%WER on eval98 using VTLN triphone/quinphone models and a trigram LM
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

**MMIE model training**

- More recent work presented on Tuesday (Povey and Woodland)
- Lattice based MMIE using the Extended Baum-Welch algorithm
- Initialisation from h5train00 MLE models
- Lattices on training data using MLE models and a bigram language model
- Optimal performance after 2 iterations

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Swbd-II</th>
<th>CHE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>42.5</td>
<td>47.7</td>
<td>45.1</td>
</tr>
<tr>
<td>1</td>
<td>40.7</td>
<td>46.2</td>
<td>43.5</td>
</tr>
<tr>
<td>2</td>
<td>40.3</td>
<td>45.1</td>
<td>42.7</td>
</tr>
</tbody>
</table>

%WER on eval98 using VTLN triphone models and a trigram language model.
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

**Pronunciation probabilities**

- Unigram probabilities scaled by LM weight
- Including silence models at word ends
- Estimated from alignments on h5train00
  - Smoothing of unseen pronunciation variants
  - Normalisation
- Marginal difference if estimated on alignments with different model sets

<table>
<thead>
<tr>
<th>System Type</th>
<th>Triphones</th>
<th></th>
<th>Quinphones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd-II</td>
<td>CHE</td>
<td>Total</td>
</tr>
<tr>
<td>ST/GD</td>
<td>41.4</td>
<td>47.0</td>
<td>44.2</td>
</tr>
<tr>
<td>ST/GD/PP</td>
<td>40.1</td>
<td>45.5</td>
<td>42.8</td>
</tr>
<tr>
<td></td>
<td>Swbd-II</td>
<td>CHE</td>
<td>Total</td>
</tr>
<tr>
<td>ST/GD</td>
<td>41.0</td>
<td>46.1</td>
<td>43.6</td>
</tr>
<tr>
<td>ST/GD/PP</td>
<td>39.2</td>
<td>44.6</td>
<td>41.9</td>
</tr>
</tbody>
</table>

%WER on eval98 using VTLN GI triphone/quinphone models
Full-Variance transform

- Block diagonal transformation matrix $H$ (Gales 1997)

$$\hat{\Sigma} = H\Sigma H^T$$

- Implemented efficiently by transforming means + data

- Speaker dependent semi-tied covariance

- Single global transform used after MLLR
Confusion Networks

- Estimation of word posterior probabilities (Mangu et al.)
  - Link posterior probability estimation in lattices
  - Clustering to obtain a linear graph

- Used for
  - Minimum word error rate decoding (Stolcke et al.)
    → Selection of word with maximum posterior from each confusion set
  - Confidence scores
  - System combination

- Consistent improvement of approx. 1% WER absolute

- Confusion network combination (CNC) gives 0.3–0.4% absolute WER improvement over ROVER (using confidence scores)
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

**System Results (1)**

<table>
<thead>
<tr>
<th></th>
<th>eval98</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd-II</td>
<td>CHE</td>
<td>Total</td>
</tr>
<tr>
<td>P1</td>
<td>47.0</td>
<td>51.6</td>
<td>49.3</td>
</tr>
<tr>
<td>P2</td>
<td>40.0</td>
<td>44.9</td>
<td>42.5</td>
</tr>
<tr>
<td>P3</td>
<td>37.5</td>
<td>42.4</td>
<td>40.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>eval00</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd-I</td>
<td>CHE</td>
<td>Total</td>
</tr>
<tr>
<td>P1</td>
<td>31.7</td>
<td>45.4</td>
<td>38.6</td>
</tr>
<tr>
<td>P2</td>
<td>25.5</td>
<td>38.1</td>
<td>31.8</td>
</tr>
<tr>
<td>P3</td>
<td>22.9</td>
<td>35.7</td>
<td>29.3</td>
</tr>
</tbody>
</table>

P1
GI, MLE triphones, 27k, trigram
VTLN warp factors, gender

P2
GI, MMIE triphones, 54k, 4-gram

P3
MLLR
GI, MMIE triphones, 54k, bigram
4-gram Lattices
New Features in the CU-HTK System for Transcription of Conversational Telephone Speech

**System Results (2)**

![Diagram of system results](image)

- **GI, MMIE**
  - P4a
  - P5a
  - P6a
  - 2 MLLR Trans.

- **GD, MLE, ST**
  - P4b
  - P5b
  - 4-gram Lattices

- **MLLR**
  - FV
  - PPROB
  - CN

- **Final result**
  - 1-best CN Lattice

---

**Legend**
- Lattice
- CN
- 1-best
### System Results(3)

<table>
<thead>
<tr>
<th></th>
<th>eval98</th>
<th></th>
<th>eval00</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd-II</td>
<td>CHE</td>
<td>Total</td>
<td>Swbd-I</td>
</tr>
<tr>
<td>P4a</td>
<td>34.5</td>
<td>39.6</td>
<td>37.1</td>
<td>20.9</td>
</tr>
<tr>
<td>P4b</td>
<td>35.5</td>
<td>40.3</td>
<td>37.9</td>
<td>21.9</td>
</tr>
<tr>
<td>P5a</td>
<td>33.9</td>
<td>38.4</td>
<td>36.2</td>
<td>20.3</td>
</tr>
<tr>
<td>P5b</td>
<td>34.5</td>
<td>39.5</td>
<td>37.0</td>
<td>21.0</td>
</tr>
<tr>
<td>P6a</td>
<td>33.6</td>
<td>38.4</td>
<td>36.0</td>
<td>20.3</td>
</tr>
<tr>
<td>CNC</td>
<td>32.5</td>
<td>37.4</td>
<td>35.0</td>
<td>19.3</td>
</tr>
</tbody>
</table>

% WER on eval98 and eval00
Conclusions

• Improvements
  – MMIE 2.5–3.0%
  – Soft-tying 0.5–1.0%
  – Pronunciation probabilities 1.0-1.5%
  – Full variance transform 0.5-0.7%
  – Confusion network decoding 1.0%
  – Confusion network combination 0.3%
  – Better Confidence scores

• Overall reduction in word error rate by 11% relative

• Best system performance in NIST Hub5 English 2000 (and 2001) evaluations
Update: The CU-HTK 2001 system

- Recent evaluation in March 2001

- Test sets (dev/eval)
  - Swbd-I, Swbd-II, Swbd cellular

- Improved system
  - Resegmentation for CMN/CVN/VTN computation
  - Improved MMI training
  - Lattice based iterative MLLR

<table>
<thead>
<tr>
<th></th>
<th>Swbd-I</th>
<th>Swbd-II</th>
<th>Swbd cell.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 2000</td>
<td>19.3</td>
<td>32.5</td>
<td>33.2</td>
<td>28.3</td>
</tr>
<tr>
<td>System 2001</td>
<td>18.3</td>
<td>31.9</td>
<td>32.1</td>
<td>27.3</td>
</tr>
</tbody>
</table>

% WER on the 2001 official development set
• Hidden Markov Model Toolkit available for free (http://htk.eng.cam.ac.uk)

• Meeting for people using or interested in HTK

• Meet developers and other users of HTK3

• Discuss new version and future development

Today 18:30 Hilton Salon 2