The Johns Hopkins University 2002
Large Vocabulary Conversational Speech Recognition System

B. Byrne, V. Doumpiotis, S. Kumar, S. Tsakalidis, V. Venkataramani
Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University
Baltimore, MD 21218, USA
CLSP Evaluation System

Our goals in entering the evaluation:

- Develop a reasonably good system for the development of new LVCSR techniques
- Incorporate new acoustic modeling and decoding techniques
- Provide experience for CLSP students

The evaluation system was built and run entirely by graduate students:

- Shankar Kumar - MBR decoding
- Veera Venkataramani - Acoustic Processing
- Vlasios Doumiotis - Discriminative Speaker Adaptive Training
- Stavros Tsakalidis - Discriminative Likelihood Linear Transforms

Effort was mainly on the incorporation of new research results
- relatively little system optimization or tuning
What’s New for 2002

- What’s Fixed
  - Parameterized Cepstral Transforms for VTN with mean and variance normalization
  - Corrected language model back-off in lattice rescoring
  - Pronunciation probabilities
  - Multi-System Rescoring

- What’s New For Us
  - MMI

- What’s New: MMI with CMLLR-Based Acoustic Normalization
  - Discriminative Speaker Adaptive Training - DSAT
  - Discriminative Likelihood Linear Transforms - DLLT
  - SMBR Lattice Combination and SMBR Decoding

- What’s History
  - Unsupervised Discriminative CMLLR Adaptation ... for the moment
  - Maximum Likelihood Speaker Adaptive Training ... replaced by DSAT
System Overview

Primary System
Stage 1: Initial Transcription / Channel and Vocal Tract Normalization
Stage 2: MLLR Speaker Adaptation / Lattice Generation
Stage 3: Lattice Rebuilding
Stage 4: Segmental Minimum Bayes-Risk (SMBR) Decoding

Contrast System
Stage 5: Lattice Generation with DLLT+MMI Acoustic Models
Stage 6: Lattice Generation with DSAT+MMI Acoustic Models
Stage 7: MultiSystem SMBR Decoding

Thanks to ...
Michael Riley - AT&T Large Vocabulary Decoder and FSM Toolkit
Murat Saraclar - Advice in using the AT&T FSM Tools for discriminative training
Andreas Stolcke - SRI LVCSR trigram language models and SRI LM Toolkit
CUED - HTK Hidden Markov Toolkit
Acoustic Processing

PLP cepstral features
- First and second derivatives
- Conversation-side cepstral mean normalization

Vocal Tract Normalization
- Sine-Log All-Pass Transforms (SLAPT-2)
- Based on Ph.D. work of J. McDonough
- Gaussian Mixture Models to find warps
- VTN performed prior to initial decoding
- Amoeba Search for optimal parameter values

Conversation Side Variance Normalization
- separate normalization for speech and silence

HNORM: HTK Library Module
Implements linear VTN, BLT, and SLAPT with various search strategies as well as CSM+VN
Pronunciation Lexicon

Pronlex (LDC) based pronunciations
- 42 context independent phones, with silence and short pause models
- augmented by Mississippi State transcriptions

Pronunciations are marked to indicate:
- word boundary information
- monosyllabic words
- interjections

Multiword models taken from SRI 2000 LVCSR system
- courtesy of Andreas Stolcke
- word bigrams and trigrams that occurred more than 200 times
- standard pronunciations

```
ad
abalone ae:s b ax l ow n iy:e
ah ahI
all-around ao:s l:e ax:s r aw n d:e
ah-ya-ya-ya ahI y:s aa:e y:s aa:e y:s aa:e
```

Pronunciation probabilities estimated over the acoustic training data
Primary System: Acoustic and Language Models

Acoustic Models

Stage 1: For initial transcription for VTN and variance normalization
- Standard HTK flat-start training procedure
- Tied state, cross-word, context-dependent triphones
- 8340 unique triphone states
- 16 mixtures per speech state
- tagged acoustic clustering to incorporate interjection and word-boundary info

Stage 2: SLAPT-2 Vocal Tract Normalization and MMI
- Stage 1 Models were retrained using SLAPT-2 normalized acoustic training data
- Further refined by three iterations of Viterbi-style MMI over triphone lattices generated in the usual way

Language Models

SRI 2000 LVCSR Evaluation Multiword bigram LM, courtesy of Andreas Stolcke
Interpolated trigrams LMs : Broadcast News (130M), Switchboard (3M), CallHome (210K)
- 33K word vocabulary
- 417325 bigrams and 200696 trigrams
- heavily pruned for Stage 1 decoding
Contrast System: Discriminative Linear Transforms

Goal: Discriminative versions of Maximum Likelihood training procedures

Techniques that incorporate ML estimation of linear transforms for acoustic normalization and adaptation:
- Maximum Likelihood Linear Transforms (MLLT)
- Speaker Adaptive Training (SAT)

Linear transforms of HMM parameters are estimated along with HMM Gaussian means and variances
- MLLT: transform acoustic data to ease diagonal covariance Gaussian modeling assumption
- SAT: apply speaker dependent transforms to speaker independent models

Using ML transforms with MMI:
- AT&T has used CMA (similar to MLLT) followed by MMI
- McDonough (ICASSP’02) has used ML-SAT followed by MMI

Conditional Maximum Likelihood estimation procedures are available for linear transforms
- Replaces ML transform estimation
- CMLLR developed by Asela Gunawardana
- Used for Unsupervised Discriminative Adaptation in the CLSP 2001 evaluation system

CML uses an auxiliary function similar to EM
- CML versions of MLLT and SAT are readily obtained as variants of the ML algorithms
DLTs in Acoustic Training

Two different uses for Discriminative Linear Transforms during training

**Discriminative Likelihood Linear Transforms**

\[
q(o|s; \theta) = \frac{|T_r|}{\sqrt{(2\pi)^n|\Sigma_s|}} e^{-\left((T_r o - \mu_s)'\Sigma_s^{-1}(T_r o - \mu_s)/2\right)} \quad \forall s : R(s) = r
\]

Observations are transformed prior to likelihood evaluation

**Discriminative Speaker Adaptive Training**

\[
q(o|s; \theta, k) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_s|}} e^{-\left((o - T^k \mu_s)'\Sigma_s^{-1}(o - T^k \mu_s)/2\right)}
\]

Speaker dependent transforms \( T^k \) transform parameters of the observation distribution

Goal is to maximize \( P(W_i|O_i; \theta) \) by alternately updating the transforms and HMM parameters
- Both transforms and HMM parameters are estimated discriminatively
### DLLT MiniTrain / MiniTest Experiments - WER(%) 

All experiments were performed with:
- **MLLT**: ML updating of Gaussian means and variances
- **DLLT**: MMI updating of Gaussian means and variances

#### Experiments Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Swbd1</td>
<td>Swbd2</td>
<td>Swbd1</td>
</tr>
<tr>
<td><strong>Transform Only (Mean&amp;Var. Fixed)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>41.1</td>
<td>51.1</td>
<td>41.1</td>
</tr>
<tr>
<td>ML+MLLT-1it</td>
<td>39.1</td>
<td>50.3</td>
<td>38.2</td>
</tr>
<tr>
<td>ML+MLLT-2it</td>
<td>39.4</td>
<td>50.4</td>
<td>37.3</td>
</tr>
<tr>
<td>ML+DLLT-1it</td>
<td>38.5</td>
<td>49.7</td>
<td>37.8</td>
</tr>
<tr>
<td>ML+DLLT-2it</td>
<td>38.3</td>
<td>49.9</td>
<td></td>
</tr>
</tbody>
</table>

**Observations:**
- DLLT in isolation is better than MLLT (A)
- DLLT works best when initialized by MLLT (B vs. C)
DSAT Experiments - WER(%) 

All experiments were performed with MLLR rescoring with 2 regression classes
- ML-SA T : ML updating of Gaussian means and variances
- DSAT : MMI updating of only Gaussian means

<table>
<thead>
<tr>
<th></th>
<th>MiniTrain / MiniTest</th>
<th>FullSystem / Dev01</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWB1</td>
<td>SWB2</td>
</tr>
<tr>
<td>MMIE-3it</td>
<td>35.9</td>
<td>47.0</td>
</tr>
<tr>
<td>MMIE + DSAT-1it</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following were performed after the evaluation.

<table>
<thead>
<tr>
<th></th>
<th>MiniTrain / MiniTest</th>
<th>FullSystem / Dev01</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWB1</td>
<td>SWB2</td>
</tr>
<tr>
<td>MMIE + ML-SAT-1it</td>
<td>35.8</td>
<td>45.9</td>
</tr>
<tr>
<td>MMIE + ML-SAT-3it</td>
<td>35.2</td>
<td>45.2</td>
</tr>
<tr>
<td>MMIE + ML-SAT-5it</td>
<td>34.5</td>
<td>45.1</td>
</tr>
<tr>
<td>MMIE + ML-SAT + DSAT-1it</td>
<td>34.2</td>
<td>44.9</td>
</tr>
<tr>
<td>MMIE + ML-SAT + DSAT-2it</td>
<td>33.8</td>
<td>44.5</td>
</tr>
<tr>
<td>MMIE + ML-SAT + DSAT-3it</td>
<td>33.8</td>
<td>44.3</td>
</tr>
</tbody>
</table>

DSAT works best when seeded by ML-SAT
MMIE+DSAT-1it HMMs were used in the contrast system
Primary System: Segmental Minimum Bayes-Risk Decoding

Minimum Bayes-Risk (MBR) Decoder

\[ \hat{W} = \arg\min_{W' \in \mathcal{L}} \sum_{W \in \mathcal{L}} l(W, W') P(W | A) \]

Goal: Simplify a very large search problem into a sequence of smaller problems

Segmental MBR (SMBR)
- If \( l(W, W') = \sum_{i=1}^{N} l(W_i, W'_i) \), can perform MBR decoding on each sublattice
- Requires segmenting every path in the lattice wrt every other path
- Retains acoustic and language model scores from the original lattice

Approximation: Segment lattice paths wrt the MAP hypothesis \( \tilde{W} : l(\tilde{W}, W') = \sum_{i=1}^{N} l(\tilde{W}_i, W'_i) \).
- ML approximation to full MBR (ICSLP’02 submission)
Contrast System: Multi-System SMBR Decoding

Dev01 WER(%)  

<table>
<thead>
<tr>
<th></th>
<th>ML Decoding</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWB1</td>
<td>SWB2</td>
<td>SWB2C</td>
<td></td>
</tr>
<tr>
<td>MMIE</td>
<td>24.5</td>
<td>39.2</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>DLLT</td>
<td>24.0</td>
<td>38.7</td>
<td>38.8</td>
<td></td>
</tr>
<tr>
<td>DSAT</td>
<td>24.5</td>
<td>39.3</td>
<td>39.5</td>
<td></td>
</tr>
<tr>
<td>Single System SMBR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMIE + SMBR</td>
<td>24.0</td>
<td>38.3</td>
<td>38.8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MulitSystem SMBR</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lattice Intersection</td>
<td>24.0</td>
<td>38.4</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>SMBR-Intersection</td>
<td>23.5</td>
<td>37.8</td>
<td>38.0</td>
</tr>
<tr>
<td></td>
<td>SMBR-Union</td>
<td>23.3</td>
<td>37.8</td>
<td>37.8</td>
</tr>
</tbody>
</table>

Observations:
- SMBR is better than simply intersecting lattices and rescoring
- Adding posteriors (SMBR-Union) over sublattices is better than multiplying them (SMBR-Intersection)
## Performance of Primary and Contrast Systems

<table>
<thead>
<tr>
<th></th>
<th>Dev01</th>
<th>Eval02</th>
<th>xRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SWB1</td>
<td>SWB2</td>
<td>SWB2C</td>
</tr>
<tr>
<td><strong>Primary System</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSMVN+VTN</td>
<td>29.4</td>
<td>44.6</td>
<td>44.8</td>
</tr>
<tr>
<td>First Pass: CSMVN+VTN</td>
<td>25.7</td>
<td>41.0</td>
<td>40.9</td>
</tr>
<tr>
<td>Lattice Generation: MMI+MLLR</td>
<td>24.5</td>
<td>39.2</td>
<td>39.6</td>
</tr>
<tr>
<td>Lattice LM Rebuilding+Rescoring</td>
<td>24.0</td>
<td>38.3</td>
<td>38.8</td>
</tr>
<tr>
<td>SMBR Decoding</td>
<td>25.2</td>
<td>40.0</td>
<td>39.7</td>
</tr>
<tr>
<td><strong>Contrast System</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lattice Rescoring: DLLT</td>
<td>24.0</td>
<td>38.7</td>
<td>38.8</td>
</tr>
<tr>
<td>Lattice Rescoring: DSAT</td>
<td>24.5</td>
<td>39.3</td>
<td>39.5</td>
</tr>
<tr>
<td>Lattice Combination+SMBR Decoding</td>
<td>23.3</td>
<td>37.8</td>
<td>37.8</td>
</tr>
</tbody>
</table>

Speed measured on dual cpu 1.2GHz Athon processors with 1GB RAM
Summary

New Developments in the CLSP 2002 Evaluation System

- MMI with CMLLR-Based Acoustic Normalization
  - Discriminative Speaker Adaptive Training - DSAT
  - Discriminative Likelihood Linear Transforms - DLLT
  - Nice gains from DSAT, unfortunately not in time for the evaluation

- Multi-System SMBR Decoding

- HTK Library module for acoustic normalization

For more info, see our evaluation website:

http://www.clsp.jhu.edu/research/rteval