Approaches in Very Large Vocabulary Continues Speech Recognition

Afeka Center for Language Processing (ACLP)

Dikla Shchnike
Speech Recognition Day – 08.06.10
Motivation

The Need:
- LVCSR engine target market applications:
  - Dictation
  - VM Transcription
  - Navigation
  - Call Centers...

Main Challenges:
- Performance
- Computational complexity
- RT in large scale deployment
Research Focus

Performance vs. Computational Complexity using various approaches
Classic LVCSR System Overview

Training Speech Data → Acoustic Models (AM) → Textual Output

Training

Speech Input

Language Model (LM) Estimation

Decoding

LM
Lexicon
The Training Phase: **AM**

- **Transcription**
  - Text Files
  - Wave Files

- **Lexicon**
  - Cat Eat ..... K AE T
  - Eat IY T ..... K AE T IY T ...

- **Feature Extraction (FE)**
  - Framing
  - Feature Vector

- **Training Process**
  - Segmental K-Means

- **Acoustic Models**
- Large textual corpora
- Variety of domains
- N-Gram:
  - Uni-gram
  - Bi-gram
  - Tri-gram
  - ...

The Training Phase: $LM$
The Decoding Phase

Speech Input

Feature Extraction

Grid Search
(Decoding)

Textual Output

Built Grid

start

end

Word 1

Word 2

Word N

LM

Lexicon

Acoustical Models
Classic LVCSR: *Main Advantages*

- One-time grid construction
- One stage of decoding
- Focused search
Classic LVCSR: Disadvantages

- High computational complexity
- Huge grid size
- Response time
- Mediocre recognition performance
- Every word must be included in the LM
Experimental Environment

- Language: American English
- Database: Macrophone
  - Channel: Telephone
  - No. of speakers: 4505
  - No. of recordings per speaker: 44
  - 1,037,924 total words (tokens)
  - 12,092 unique words (types)
- Read Speech
Experimental Environment: \textit{Train}

- **Database:** Macrophone Training (4005 speakers)
- **Features:** MFCC $13 + \Delta + \Delta\Delta$
- **Acoustic Model Topology:**
  - HMM 3 state left to right
  - Tied state triphones
  - 16 Mix
- **LM Topology:**
  - Bi-Gram
Experimental Environment: *Test*

- DB: Macrophone DEVTEST (500 Speakers)
- Sentences only
- 3139 Utterances
- Lexicon size: 8.7K
Performance Measures

For phoneme, word and character sequences

- Correct Rate

\[ CR = \frac{Hit}{Total\ Words} \cdot 100 \]

- Error Rate

\[ ER = \frac{Deletion + Insertion + Subsition}{Total\ Words} \cdot 100 \]

- Recognition Rate

\[ RR = 100 - ER \]
Performance vs. Sequence Length

- Character Correct Rate
- Character Recognition Rate
- Word Correct Rate
- Word Recognition Rate
Performance vs. Vocabulary Size – 10 word Utterance

- Word Correct Rate
- Character Correct Rate
- Character Recognition Rate
- Word Recognition Rate

Lexicon Size

40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90%

400 8700 54400
Multi-Stage LVCSR

Grammar

Bi-phone

Phoneme Decoder

DH A I A B N M T P ……

Speech Input

AM

Search Grid (Decoding)

Built Grid

Lexicon

Vocabulary Reduction

Reduce Voc.

Speech Input

Textual Output

LM
Multi-Stage LVCSR: 
Potential Advantages

- Increased performance
  - The LVCSR works with a reduced vocabulary
- Reduced computational complexity
  - Additional stage/s vs. smaller LVCSR search space
Multi Stage LVCSR: Disadvantages

- Phoneme decoder recognition rate.
- Highly dependent on second stage coverage rate
- Sequential – potential increase in overall error rate
## Classic vs. Multi stage

### First Stage - Phoneme Decoder

<table>
<thead>
<tr>
<th>PCR [%]</th>
<th>PER [%]</th>
<th>PRR [%]</th>
<th>Processing Time [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.97</td>
<td>56.87</td>
<td>43.13</td>
<td>3.3</td>
</tr>
</tbody>
</table>

### Second Stage – Vocabulary Reduction

<table>
<thead>
<tr>
<th>Lexicon Size</th>
<th>Reduced Voc.</th>
<th>Coverage [%]</th>
<th>Processing Time [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.7K</td>
<td>4.4K</td>
<td>93.38%</td>
<td>0.32</td>
</tr>
<tr>
<td>54.4K</td>
<td>27K</td>
<td>96.08%</td>
<td>2.25</td>
</tr>
</tbody>
</table>

### Classic LVCSR vs. Third Stage - Multi stage LVCSR

<table>
<thead>
<tr>
<th>Lexicon Size</th>
<th>Classic LVCSR</th>
<th>Third Stage - Multi stage LVCSR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCR [%]</td>
<td>CER [%]</td>
</tr>
<tr>
<td>8.7K</td>
<td>75.29</td>
<td>33.35</td>
</tr>
<tr>
<td>54.4K</td>
<td>75.48</td>
<td>35.00</td>
</tr>
</tbody>
</table>
Next Steps

- A more extensive set of experiments should be conducted
  - Test on spontaneous speech
  - Increased lexicon size – 100k, 250k
- Test LVCSR with Pruning
- Lattice – in phoneme decoder stage
- Replace grid search with textual algorithm
The End
Acoustic Model Topology

- 39 monophones
- 6 noises
- #Triphones before DCT: 7865
- #Triphones after DCT: 3377
- #States After DCT: 1476
- Global covariance
- 16 mixtures for triphones
- 32 mixtures for noise.
### Test Set Information

- Sentences from: Timit, ATIS, WSJ

<table>
<thead>
<tr>
<th>Item</th>
<th>#Utt.</th>
<th>Average Phoneme in Sentence</th>
<th>Average Word in Sentence</th>
<th>Sentence Duration [sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>1334</td>
<td>36.19</td>
<td>7.9</td>
<td>3.31</td>
</tr>
<tr>
<td>ATIS</td>
<td>925</td>
<td>37.49</td>
<td>9.38</td>
<td>3.48</td>
</tr>
<tr>
<td>WSJ</td>
<td>880</td>
<td>39.06</td>
<td>9.08</td>
<td>3.61</td>
</tr>
</tbody>
</table>
## Phoneme Decoder Performance

<table>
<thead>
<tr>
<th>Item</th>
<th>#Utt.</th>
<th>Correct [%]</th>
<th>Del [%]</th>
<th>Sub [%]</th>
<th>Ins [%]</th>
<th>PER [%]</th>
<th>PRR [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timit</td>
<td>1334</td>
<td>58.19</td>
<td>7.64</td>
<td>34.16</td>
<td>17.11</td>
<td>58.92</td>
<td>41.08</td>
</tr>
<tr>
<td>WSJ</td>
<td>880</td>
<td>58.31</td>
<td>7.78</td>
<td>33.92</td>
<td>16.37</td>
<td>58.06</td>
<td>41.94</td>
</tr>
<tr>
<td>ATIS</td>
<td>925</td>
<td>64.12</td>
<td>6.88</td>
<td>29.00</td>
<td>16.93</td>
<td>52.81</td>
<td>47.19</td>
</tr>
<tr>
<td>VM1</td>
<td>1341</td>
<td>46.29</td>
<td>14.40</td>
<td>39.31</td>
<td>23.69</td>
<td>77.39</td>
<td>22.61</td>
</tr>
<tr>
<td>VM2</td>
<td>1144</td>
<td>45.56</td>
<td>14.39</td>
<td>40.04</td>
<td>26.32</td>
<td>80.76</td>
<td>19.24</td>
</tr>
</tbody>
</table>
Processing Time Information

Full Analysis of 3 Utt.

- Feature Extraction – 0 sec
- Built Grid

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>47 nodes 1669 arcs</td>
<td>4403 nodes 29803 arcs</td>
</tr>
<tr>
<td>29803 arcs</td>
<td>8759 nodes 60537 arcs</td>
</tr>
<tr>
<td>60537 arcs</td>
<td>27228 nodes 78027 arcs</td>
</tr>
<tr>
<td>78027 arcs</td>
<td>54451 nodes 151922 arcs</td>
</tr>
<tr>
<td>151922 arcs</td>
<td>8756 nodes 6426195 arcs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 [sec]</td>
<td>0 [sec]</td>
</tr>
<tr>
<td>1 [sec]</td>
<td>1 [sec]</td>
</tr>
<tr>
<td>1 [sec]</td>
<td>2 [sec]</td>
</tr>
<tr>
<td>2 [sec]</td>
<td>37 [sec]</td>
</tr>
</tbody>
</table>

- Decoding

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>47 nodes 1669 arcs</td>
<td>4403 nodes 29803 arcs</td>
</tr>
<tr>
<td>29803 arcs</td>
<td>8759 nodes 60537 arcs</td>
</tr>
<tr>
<td>60537 arcs</td>
<td>27228 nodes 78027 arcs</td>
</tr>
<tr>
<td>78027 arcs</td>
<td>54451 nodes 151922 arcs</td>
</tr>
<tr>
<td>151922 arcs</td>
<td>8756 nodes 6426195 arcs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Bigram</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3 [sec]</td>
<td>7 [sec]</td>
</tr>
<tr>
<td>13 [sec]</td>
<td>31 [sec]</td>
</tr>
<tr>
<td>31 [sec]</td>
<td>65 [sec]</td>
</tr>
<tr>
<td>65 [sec]</td>
<td>83 [sec]</td>
</tr>
</tbody>
</table>
Computer Parameters

- Intel Dual Core CPU
- Processor Clock 2.5GHz
- 3GB of RAM