Extracting Posteriors from a Gaussian Mixture Model

Xinkun Nie and Paul Myers
6.375 Final Project
May 11, 2016
Introduction

Voice Activity Detection

Raw Speech Vectors

Posterior Extraction from GMMs

Dimensionality Reduction

Applications

Classification

$P(\theta \mid \text{Data})$
System Overview

- All computations are performed on 32-bit fixed-point numbers
- GMMs and frame matrix are stored in DRAM
Memory Layout

DRAM

Diagonal GMM
- G-Constant
- Inverse Variance
- Mean * Inverse Variance

Full GMM
- G-Constant
- Mean * Inverse Covariance
- Inverse Covariance

Frames
- Raw Frame
- Frame\(^2\)
- Frame * Frame\(^T\)
Diagonal GMM Selection Module

- Computes two dot products
- 60 elements, 32 bits
Sorting Module

- Chooses $K$ largest log-likelihoods
**Full GMM Selection Module**

- Computes two dot products
- 60 elements, 32 bits
- 1830 elements, 32 bits
Simulation Results

<table>
<thead>
<tr>
<th>Gaussian ID</th>
<th>Hardware Value</th>
<th>Software Value</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-156.495</td>
<td>-156.5087</td>
<td>0.00875</td>
</tr>
<tr>
<td>4</td>
<td>-161.275</td>
<td>-161.2985</td>
<td>0.0145</td>
</tr>
<tr>
<td>5</td>
<td>-195.919</td>
<td>-195.9338</td>
<td>0.00755</td>
</tr>
<tr>
<td>6</td>
<td>-181.110</td>
<td>-181.1254</td>
<td>0.00850</td>
</tr>
<tr>
<td>8</td>
<td>-192.481</td>
<td>-192.4866</td>
<td>0.00290</td>
</tr>
</tbody>
</table>

- Setup: 10 Gaussians, 2 frames, 5 selected
- Average error for the frame shown is 0.0085%
## Synthesis Results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slice LUTs</td>
<td>295450 (97.31% utilization)</td>
</tr>
<tr>
<td>Slice Registers</td>
<td>260133 (42.84% utilization)</td>
</tr>
<tr>
<td>Block RAM Tiles</td>
<td>354 (34.36% utilization)</td>
</tr>
<tr>
<td>DSP Blocks</td>
<td>1920 (68.57% utilization)</td>
</tr>
<tr>
<td>Clock Frequency</td>
<td>20 MHz</td>
</tr>
<tr>
<td>Critical Path</td>
<td>33.405 ns (In DRAM Control)</td>
</tr>
<tr>
<td>Total Negative Slack</td>
<td>0.000 ns</td>
</tr>
<tr>
<td>Total Negative Slack Failing Endpoints</td>
<td>277973</td>
</tr>
<tr>
<td>Worst Negative Slack</td>
<td>0.077 ns</td>
</tr>
</tbody>
</table>

- Area was a significant concern
- Routing difficulties may have led to timing violations
- Could have increased BRAM utilization
Questions?