GPU Accelerated Model Combination for Robust Speech Recognition and Keyword Search

Wonkyum Lee  Jungsuk Kim  Ian Lane
Electrical and Computer Engineering
Carnegie Mellon University

March 26, 2014 @GTC2014
Overview

- Introduction
- Acoustic Model
- Acoustic Model combination
- GPU Accelerated Model Combination
- Evaluation Results
- Summary
Introduction
Introduction

ASR (Automatic Speech Recognition)

- Feature Extraction
- Acoustic Model
- Lexicon
- Language Model
- Decoder
- Word String
- Word #1, Word #2, ...
Introduction

KWS (Keyword Search)

Speech-To-Text (by ASR)
Welcome to GTC two thousands fourteen

Indexer

Keyword Search Task
Keyword:
GTC

Thousands of Hours of Indexed Audio

Hit:
GTC
Introduction

• Speech Recognition
  ▪ Speech to Text
  ▪ Evaluation Metric: Word Error Rate

• KWS
  ▪ Spot the Keyword in Audio
  ▪ Evaluation Metric: Actual Term Weighted Value

- > *Both Tasks require Robust ASR!*
Acoustic Model

Input Acoustic Signal

Input Transcript

Recognize

Speech

Output Trained Acoustic Model
Acoustic Model

Gaussian Mixture Model (GMM)
Acoustic Model

• GMM/HMM
  ▪ GMMs trained using the EM algorithm are able to self organize to fit a data set
  ▪ Hidden Markov Model models sequential patterns
  ▪ Technical Advances over past 10 years
    • Adaptation, Discriminative Training, SGMM
Acoustic Model

Deep Neural Network (DNN)

George E. Dahl, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”
Acoustic Model

• DNN/HMM
  ▪ Called “Hybrid DNN/HMM system”
  ▪ Has good discrimination
  ▪ Temporal aspects are deal with HMM, like left-to-right HMM models
  ▪ Drawback is computation is expensive!
Acoustic Model Combination

• How can we improve ASR with Acoustic Model
  ▪ Robust Acoustic Model
    • More and More Data -> Better and Better Accuracy
    • Robust Feature(Bottle-neck Feature, Noise Robust Feature)
  ▪ Acoustic Model Combination
Acoustic Model Combination

Model Structure

GMM1  DNN1

GMM2  DNN2

Log likelihood Pattern by Acoustic Model
Acoustic Model Combination

• Different Acoustic Models (model structure, features) have distinct speech recognition pattern.
  -> different performance in Speech Recognition and Keyword Search

The goal is to find a way to combine different acoustic models for robust speech recognition and keyword search

Consideration

• Data type to be combined for AM combination
• Weighting criterion
• Total system run time (Real-time factor)
Acoustic Model Combination

- **Multi-stream based AM combination**
  - Combine multiple AMs at the AM score level
  - Weighting Criterion (Arithmetic, Geometric, Harmonic)
  - *One-pass and One time decoding*

- **Other combination Method**
  - Lattice Combination, Rover, Combmnz
  - Intermediate/output level combination
Acoustic Model Combination

Features

\[ o_1 \]
\[ o_N \]

\[ AM_1 \]
\[ S_{1,1} \quad S_{1,2} \quad S_{1,3} \]
\[ S'_{1} \quad S'_{2} \quad S'_{3} \quad S'_{4} \]

\[ AM_N \]
\[ S_{N,1} \quad S_{N,2} \]
\[ S'_{1} \quad S'_{2} \quad S'_{3} \quad S'_{4} \]

\[ \sum \]
\[ w_1: \text{weight (normalization)} \]

WFST(combined)

DECODER

Words

Combination:
- Arithmetic mean
- Geometric mean
- Harmonic mean

Decoding
Acoustic Model Combination

GPU Accelerated Speech Recognition - Talked at GTC 2013

Speech recognition contains many highly parallel tasks + GPU processors optimized for parallel computing = HYDRA an ASR engine designed specifically for GPUs
Acoustic Model Combination

GMM scores
DNN scores
GMM scores
DNN scores
GMM scores
DNN scores
GMM scores
DNN scores

Speech to Text
Decoder
KWS
combined scores
Detected Terms

Computation on GPU
CPU
Experimental evaluations

• **Data:**
  • IARPA BABEL Program [Vietnamese] language collection: babel107b-v0.7[1]
  • Limited language pack (10 hrs training, 20 hrs test)

• **Features:**
  • **LMEL:** Log Mel filter bank coefficients
  • **MFCC:** Mel Frequency Cepstral Coefficients
  • **BNF:** Bottlenect features
  • **FFV:** Fundamental Frequency Variation feature
  • **Pitch:** Pitch tracking feature

<table>
<thead>
<tr>
<th>Features</th>
<th>Dim.</th>
<th>Source feature</th>
<th>Input frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNF₁</td>
<td>42</td>
<td>40&lt;sup&gt;th&lt;/sup&gt; lmel + FFV</td>
<td>11</td>
</tr>
<tr>
<td>BNF₂</td>
<td>42</td>
<td>30&lt;sup&gt;th&lt;/sup&gt; lmel + FFV + Pitch</td>
<td>11</td>
</tr>
<tr>
<td>BNF₃</td>
<td>42</td>
<td>20&lt;sup&gt;th&lt;/sup&gt; MFCC + FFV</td>
<td>11</td>
</tr>
</tbody>
</table>

## Experimental evaluations

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>Tree</th>
<th>WER (%)</th>
<th>ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM₁</td>
<td>BNF₁</td>
<td>Tree₁</td>
<td>68.0</td>
<td>0.1341</td>
</tr>
<tr>
<td>GMM₂</td>
<td>BNF₂</td>
<td></td>
<td>69.5</td>
<td>0.1271</td>
</tr>
<tr>
<td>GMM₃</td>
<td>BNF₃</td>
<td></td>
<td>71.5</td>
<td>0.1171</td>
</tr>
<tr>
<td>DNN₁</td>
<td>BNF₁</td>
<td>Tree₁</td>
<td>67.3</td>
<td>0.1377</td>
</tr>
<tr>
<td>DNN₂</td>
<td>BNF₂</td>
<td></td>
<td>68.3</td>
<td>0.1328</td>
</tr>
<tr>
<td>DNN₃</td>
<td>BNF₃</td>
<td></td>
<td>69.8</td>
<td>0.1034</td>
</tr>
</tbody>
</table>

Baseline system performance

- Trained 6 *acoustic models* (3 GMMs, 3 DNNs) with 3 different features
Experimental evaluations

<table>
<thead>
<tr>
<th>Combination scheme</th>
<th>WER</th>
<th>ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best single system (DNN₁)</td>
<td>67.3</td>
<td>0.1377</td>
</tr>
<tr>
<td>Arithmetic mean</td>
<td>63.6 (-3.7)</td>
<td>0.1794 (+30.3%)</td>
</tr>
<tr>
<td>Geometric mean</td>
<td>65.4 (-2.9)</td>
<td>0.1706 (+23.9%)</td>
</tr>
<tr>
<td>Harmonic mean</td>
<td>66.2 (-1.1)</td>
<td>0.1666 (+21.0%)</td>
</tr>
</tbody>
</table>

WER and ATWV for different combination schemes

- Combined 6 acoustic models (GMM₁+2+3+DNN₁+2+3)
- Arithmetic mean showed the most improved performance.
  - 3.7% absolute WER improvement
  - 30.3% relative ATWV improvement
Experimental evaluations

- **RTF**
- **ATWV**

<table>
<thead>
<tr>
<th></th>
<th>RTF</th>
<th>ATWV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GPU-search</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GPU-based AM computation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-Models</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental evaluations

- State-level combination obtains best WER vs. Lattice Comb., Rover
  - Note: same phone-states used across all models

- **CombMNZ** obtains better ATWV when combining more than 2 models
  - However 5x - 10x slower
  - At comparable RTF: **Multi-stream=0.23 > CombMNZ=0.20**

<table>
<thead>
<tr>
<th>Models</th>
<th>Multistream</th>
<th>CombMNZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 model</td>
<td>67.3% (0.14)</td>
<td>0.14</td>
</tr>
<tr>
<td>2 models</td>
<td>64.7% (0.17)</td>
<td>✔️ 0.16</td>
</tr>
<tr>
<td>4 models</td>
<td>63.6% (0.18)</td>
<td>✔️ 0.19</td>
</tr>
<tr>
<td>6 models</td>
<td>63.6% (0.18)</td>
<td>✔️ 0.20</td>
</tr>
<tr>
<td>6 models (large lattice)</td>
<td>64.7% (0.23)</td>
<td>✔️ &quot;</td>
</tr>
</tbody>
</table>
Conclusion

• Proposed **Multi-stream Model combination** in GPU accelerated speech recognition framework
• Multi-stream combination gives comparable performance with efficient runtime
• Future work
  • **More combination schemes:**
    • Weighted model combination (Model, HMM state level weights)
    • DNN-based combination
  • **Faster decoding speed:**
    • Use of CUDA multi-stream technique.
Thank you for your attention.