2009 NIST Language Recognition Evaluation
Evaluation Overview

Craig Greenberg
Alvin Martin

Based on the NIST presentation at:

LRE09 Workshop
Baltimore, Maryland, USA
June 24-25, 2009
Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
• Performance History
• Performance by Language
• Performance by Data Type
• Summary
What’s New For LRE09?

• Primary (new) data is broadcast telephone bandwidth Voice of America (VOA) data
  – Early analysis of VOA data done at Brno
  – Collected and audited by the LDC
  – Large VOA corpora and designated segments made available for development in languages for which previous LRE conversational telephone speech (CTS) data not available
• 23 target languages, 16 out-of-set languages
• Larger numbers of test segments available for most languages
• Segments of approximately 3, 10, or 30 seconds of speech all grouped together (but performance examined separately)
  – Careful listening to 10 and 3 second CTS segments
  – Found overlapping 10 and 3 second CTS speech segments that minimized time elapsed
  – Selected 10 and 3 second VOA by iterating over each sample and:
    • Let $E_{avg_i}$ be the average energy in candidate segment $seg_i$
    • Let $E_{max}$ be the maximum of $E_{avg_i}$ over all $seg_i$
    • Let $score_i$ be the score for segment $seg_i$, with $score_i = \max(E_{w1}, E_{w2}, .05*E_{max})/E_{avg_i}$
    • Chose the $seg_i$ that minimizes $score_i$.
  – Feather-cut voa segments using 10ms linear ramp
# LRE09 Languages
(counts are for 30-second segments)

<table>
<thead>
<tr>
<th>Lang.</th>
<th>VOA Train</th>
<th>VOA Test</th>
<th>CTS Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amharic</td>
<td>171</td>
<td>398</td>
<td>-</td>
</tr>
<tr>
<td>Bosnian</td>
<td>194</td>
<td>355</td>
<td>-</td>
</tr>
<tr>
<td>Cantonese</td>
<td>-</td>
<td>62</td>
<td>316</td>
</tr>
<tr>
<td>Creole-Haitian</td>
<td>186</td>
<td>323</td>
<td>-</td>
</tr>
<tr>
<td>Croatian</td>
<td>181</td>
<td>376</td>
<td>-</td>
</tr>
<tr>
<td>Dari</td>
<td>194</td>
<td>389</td>
<td>-</td>
</tr>
<tr>
<td>English-Am.</td>
<td>-</td>
<td>374</td>
<td>522</td>
</tr>
<tr>
<td>English-Ind.</td>
<td>-</td>
<td>-</td>
<td>574</td>
</tr>
<tr>
<td>Farsi</td>
<td>-</td>
<td>338</td>
<td>52</td>
</tr>
<tr>
<td>French</td>
<td>196</td>
<td>395</td>
<td>-</td>
</tr>
<tr>
<td>Georgian</td>
<td>142</td>
<td>399</td>
<td>-</td>
</tr>
<tr>
<td>Hausa</td>
<td>200</td>
<td>389</td>
<td>-</td>
</tr>
<tr>
<td>Hindi</td>
<td>-</td>
<td>397</td>
<td>270</td>
</tr>
<tr>
<td>Korean</td>
<td>-</td>
<td>318</td>
<td>145</td>
</tr>
<tr>
<td>Mandarin</td>
<td>-</td>
<td>390</td>
<td>625</td>
</tr>
<tr>
<td>Pashto</td>
<td>197</td>
<td>395</td>
<td>-</td>
</tr>
<tr>
<td>Portuguese</td>
<td>166</td>
<td>397</td>
<td>-</td>
</tr>
<tr>
<td>Russian</td>
<td>-</td>
<td>254</td>
<td>257</td>
</tr>
<tr>
<td>Spanish</td>
<td>-</td>
<td>385</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lang.</th>
<th>VOA Train</th>
<th>VOA Test</th>
<th>CTS Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turkish</td>
<td>194</td>
<td>394</td>
<td>-</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>194</td>
<td>388</td>
<td>-</td>
</tr>
<tr>
<td>Urdu</td>
<td>-</td>
<td>347</td>
<td>32</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>-</td>
<td>27</td>
<td>288</td>
</tr>
<tr>
<td>Arabic</td>
<td>-</td>
<td>Out-of-set</td>
<td>187</td>
</tr>
<tr>
<td>Azerbaijani</td>
<td>-</td>
<td>Out-of-set</td>
<td>366</td>
</tr>
<tr>
<td>Belorussian</td>
<td>-</td>
<td>Out-of-set</td>
<td>363</td>
</tr>
<tr>
<td>Bengali</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>-</td>
<td>Out-of-set</td>
<td>375</td>
</tr>
<tr>
<td>Italian</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Japanese</td>
<td>-</td>
<td>Out-of-set</td>
<td>180</td>
</tr>
<tr>
<td>Punjabi</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Romanian</td>
<td>-</td>
<td>Out-of-set</td>
<td>9</td>
</tr>
<tr>
<td>Shanghai-Wu</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Southern-min</td>
<td>-</td>
<td>Out-of-set</td>
<td>69</td>
</tr>
<tr>
<td>Swahili</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Tagalog</td>
<td>-</td>
<td>Out-of-set</td>
<td>84</td>
</tr>
<tr>
<td>Thai</td>
<td>-</td>
<td>Out-of-set</td>
<td>188</td>
</tr>
<tr>
<td>Tibetan</td>
<td>-</td>
<td>Out-of-set</td>
<td>-</td>
</tr>
<tr>
<td>Uzbek</td>
<td>-</td>
<td>Out-of-set</td>
<td>382</td>
</tr>
</tbody>
</table>
**Test Conditions**

- **Closed-set:** segment languages are limited to in-set languages, all (in-set) target languages
- **Open-set:** segment languages also include (undisclosed) out-of-set languages
- **Language pairs:** Segment and target languages limited to two, for each possible in-set pair
  - Thus always a single alternative hypothesis for each trial
  - Certain pairs designated as of particular interest

<table>
<thead>
<tr>
<th>Cantonese</th>
<th>Hindi -- Urdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portuguese</td>
<td>Farsi -- Dari</td>
</tr>
<tr>
<td>Creole</td>
<td>Bosnian -- Croatian</td>
</tr>
<tr>
<td>Russian</td>
<td>Engl. (American) -- Eng. (Indian)</td>
</tr>
</tbody>
</table>
System Input/Output

• Input: all trials for a test condition, consisting of all pairings of a test segment and a target language/dialect

• Output: for each trial
  – a decision (true/false)
  – a score on which the decision is based, where higher scores imply greater belief that “true” is the correct decision

• Systems were asked to specify if their scores could be interpreted as log-likelihood ratios (llr’s):

\[
\text{llr} = \ln P(\text{data | target language } i) - \ln P(\text{data | not target language } i)
\]

where ln is the natural logarithm function
Evaluation Rules

• All 41793 test segments of all durations must be processed for each target language
• Each test segment must be processed separately and without any knowledge of other test segments.
  – Normalization over multiple test segments is NOT allowed.
• Side knowledge of the sex or other characteristics of the test speaker is NOT allowed.
  – Unless obtained by automatic means.
• Listening to the evaluation data or any other experimental interaction with the data is NOT allowed before all test results have been submitted.
• Use of knowledge of the full set of target languages/dialects for each test IS allowed.
Basic Performance Measure

\[ C(L_T, L_N) = C_{\text{Miss}} \cdot P_{\text{Target}} \cdot P_{\text{Miss}}(L_T) \]
\[ + C_{\text{FA}} \cdot \left(1 - P_{\text{Target}}\right) \cdot P_{\text{FA}}(L_T, L_N) \]

where

- \( L_T \) and \( L_N \) are a target/non-target language pair
- \( C_{\text{Miss}}, C_{\text{FA}} \) and \( P_{\text{Target}} \) are application model parameters

For LRE09, the application parameters will be:

\( C_{\text{Miss}} = C_{\text{FA}} = 1 \), and
\( P_{\text{Target}} = 0.5 \)
Average Performance

\[ C_{\text{avg}} = \frac{1}{N_L} \sum_{L_T} \left\{ C_{\text{Miss}} \cdot P_{\text{Target}} \cdot P_{\text{Miss}}(L_T) \right\} + \sum_{L_N} C_{\text{FA}} \cdot P_{\text{Non-Target}} \cdot P_{\text{FA}}(L_T, L_N) \]

where

- \( N_L \) is the number of languages in the (closed-set) test
- \( L_O \) is the Out-of-Set “language”

\[ P_{\text{Out-of-Set}} = \begin{cases} 0.0 & \text{for the closed-set condition} \\ 0.2 & \text{for the open-set condition} \end{cases} \]

and

\[ P_{\text{Non-Target}} = \frac{1 - P_{\text{Target}} - P_{\text{Out-of-Set}}}{N_L - 1} \]
DET Curves

- In speaker recognition all trials are pooled to create the DET curve.

- In language recognition DET’s are computed separately for each language pair and then:
  - DET’s are averaged across all non-target languages to produce a DET for each target language.
  - DET’s for all target languages are averaged to produce an overall DET.

- The quality of calibration across languages affects the overall multi-target language DET curves:
  - This is illustrated dramatically for the language-pair case:
    - the DET’s for the two single targets should be symmetric.
    - these two DET’s should have the same EER.
    - but if the scores are not properly calibrated the combined DET will be degraded.
    - the next slide shows an example.
Russian-Ukrainian Pair Example

System-1
Language Pair Russian Ukrainian

System-2
Language Pair Russian Ukrainian
Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
• Performance History
• Performance by Language
• Performance by Data Type
• Summary
# Participating Sites/Teams (1)

<table>
<thead>
<tr>
<th>System Name</th>
<th>Site</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATVS</td>
<td>Universidad Autonoma de Madrid</td>
<td>Madrid, Spain</td>
</tr>
<tr>
<td>BUT-AGN</td>
<td>Brno University of Technology Agnitio</td>
<td>Brno, Czech Republic, Somerset West, South Africa</td>
</tr>
<tr>
<td>CASIA</td>
<td>Institute of Automation, Chinese Academy of Sciences</td>
<td>Beijing, China</td>
</tr>
<tr>
<td>CUHK</td>
<td>Chinese University of Hong Kong</td>
<td>N.T., Hong Kong</td>
</tr>
<tr>
<td>EHU</td>
<td>University of the Basque Country</td>
<td>Bizkaia, Spain</td>
</tr>
<tr>
<td>IFLY</td>
<td>iFlyTek Speech Lab, EEIS University of Science and Technology of China</td>
<td>HeFei, AnHui, China</td>
</tr>
<tr>
<td>IIR</td>
<td>Institute for Infocomm Research</td>
<td>Singapore</td>
</tr>
<tr>
<td>IOA</td>
<td>Institute of Acoustics, Chinese Academy of Sciences</td>
<td>Beijing, China</td>
</tr>
<tr>
<td>L2F</td>
<td>L2F-Spoken Language Systems Lab INESC-ID Lisboa</td>
<td>Lisbon, Portugal</td>
</tr>
<tr>
<td>LIA</td>
<td>Laboratorie Informatique D'Avignon</td>
<td>Avignon, France</td>
</tr>
<tr>
<td>System Name</td>
<td>Site</td>
<td>Location</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>----------</td>
</tr>
<tr>
<td>LIMSI</td>
<td>CNRS-LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur)</td>
<td>Orsay, France</td>
</tr>
<tr>
<td>LPT</td>
<td>Loquendo Politecnico di Torino</td>
<td>Torino, Italy</td>
</tr>
<tr>
<td>MIT</td>
<td>MIT Lincoln Laboratory</td>
<td>Lexington, MA, USA</td>
</tr>
<tr>
<td>NTUT</td>
<td>National Taipei University of Technology, Department of Electrical Engineering &amp; Graduate Institute of Computer and Communication Engineering</td>
<td>Taipei, Taiwan</td>
</tr>
<tr>
<td>THU</td>
<td>Tsinghua University Department of Electrical Engineering</td>
<td>Beijing, China</td>
</tr>
<tr>
<td>TNO</td>
<td>Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek</td>
<td>Soestenberg, The Netherlands</td>
</tr>
</tbody>
</table>
Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
• Performance History
• Performance by Language
• Performance by Data Type
• Summary
Overall Evaluation Results

See web page summary:

Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
  • Performance History
  • Performance by Language
  • Performance by Data Type
• Summary
Best System – Closed Set
2007, 2009

30sec
• Co-winners in 30 sec trials
• Performance loss in 30 sec trials compared with LRE07

10sec

3sec
• 3 sec saw better performance compared with LRE07
• Improved selection of 3 sec segments
Best System – Open Set
2007, 2009

30sec

10sec

3sec
Best Systems by Target Language
Closed-Set – 2007, 2009

30 sec
Korean
off chart!
Best System - Recognizing American English for American English/Indian English Language Pair

2007, 2009

- Improvement for all three durations

30sec

10sec

3sec
Best System - Recognizing Hindi for Hindi/Urdu Pair
2007, 2009

• Real improvement in 30sec and 10sec

• 3 sec still challenging
Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
• Performance History
• Performance by Language
• Performance by Data Type
• Summary
Closed Set Performance by Target Language

- Indian languages were challenging
  - CTS training somewhat better performance
Outline

• Evaluation Overview
• Participants
• Overall Evaluation Results
• Performance History
• Performance by Language
• Performance by Data Type
• Summary
Closed Set Performance by Data Type

- VOA and CTS performance broadly comparable
- CTS curves less linear, with better performance at high FA rates
Single Target Language Performance by Data Type (3sec)

System-1

System-2
Summary and Issues

• LRE09 was essentially successfully conducted largely utilizing narrowband broadcast speech
  – Performance on VOA was comparable to that with CTS
  – Larger numbers of test segments were included
  – But speakers were often repeated
• Some performance improvement seen compared with LRE07, particularly for shorter duration segments
• Similar (particularly mutually comprehensible) languages present performance (and auditing) challenges
• Some issues with scoring and DET curves
  – Should language pairs be emphasized?
  – Does LRE09 provide a model for future evaluations?