The Albayzin 2008 Language Recognition Evaluation

Luis J. Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, Amparo Varona

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
email: luisjavier.rodriguez@ehu.es

Odyssey 2010, Brno, Czech Republic
June 30, 2010
Outline

1. Context and motivation
2. The language detection task
3. Test conditions
4. Data
5. Organization
6. Results
   - CR-30 (mandatory condition)
   - CF-30
   - OF-30
   - Performance per target language
   - Segment length
   - Development conditions
7. Conclusions and current work
Context

- Spanish Thematic Network on Speech Technology
  (http://lorien.die.upm.es/ lapiz/rtth/)

- 5th Biennial Workshop on Speech Technology (Bilbao, November 2008)

- Albayzin system evaluations, on three topics: speech translation, speech synthesis and language recognition

- Software Technology Working Group (http://gtts.ehu.es): research interest on language recognition for spoken document retrieval applications
Motivation

- To promote collaboration between research groups from Spain and Portugal interested in language recognition

- To provide a speech database specifically designed for language recognition applications featuring the official languages in Spain as target languages

- To measure the accuracy that state-of-the-art systems can attain for the task of recognizing four target languages that have been in close contact from long time ago: *will this task be more challenging than expected?*

- To measure the performance of systems developed on a limited amount of data
The language detection task

As for NIST LRE: *given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment.*

**Trial:** audio segment + target language + set of non-target languages

**System output:** hard decision + score (maybe LLR)
Test conditions

- **System development**
  - Free (F): any available materials
  - Restricted (R): only those materials provided in Albayzin 2008 LRE, external data allowed neither directly nor indirectly (e.g. acoustic models in phone decoders)

- **Set of trials**
  - Closed-set tests (C): only trials corresponding to audio segments containing target languages
  - Open-set tests (O): all the trials

- **Nominal duration of audio segments**: 30, 10 and 3 seconds

- **Performance measures** (as defined in NIST LRE, using NIST software, see paper for details):
  - $C_{avg}$ ($P_{target} = 0.5, C_{miss} = C_{fa} = 1$)
  - $C_{LLR}$
  - DET curves
Database features

- Name: KALAKA (see paper at LREC 2010 for details)
- Four target languages: Spanish, Catalan, Basque and Galician
- Other languages (just to allow open-set tests): French, Portuguese, German and English
- Audio files: 16 kHz, single channel, 16 bits/sample, uncompressed PCM (WAV)
- Speech signals extracted from TV shows, including both planned and spontaneous speech in diverse environment conditions involving a varying number of speakers.
- Disjoint subsets of TV shows assigned to train, development and evaluation
- Size: around 50 hours (distributed in 3 DVD)
  - Train dataset: 36 hours (9 hours per target language)
  - Development dataset: 7,7 hours
  - Evaluation dataset: 7,7 hours
Database design issues (I)

- Only high SNR speech: fragments containing medium-high level noise, music, speech overlaps, etc. filtered out

- Segments for training had no length restrictions

- Segments for development and evaluation:
  - enclosed by a certain amount of low-energy frames
  - 3-second subset $\subset$ 10-second subset $\subset$ 30-second subset
  - length tolerance: 3-5, 10-12 and 30-33 seconds

- Development dataset (same structure for evaluation):
  - Total: 1800 segments
  - 600 segments per duration
  - 120 segments per target language and duration
  - 120 segments of unknown languages per duration
Database design issues (II)

- Proportions of unknown languages made deliberately different for development and evaluation, to avoid tuning systems to reject specific languages.
- Proportion of French and Portuguese twice the proportion of German and English.

<table>
<thead>
<tr>
<th></th>
<th>French</th>
<th>Portuguese</th>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devel</td>
<td>70</td>
<td>10</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Eval</td>
<td>10</td>
<td>70</td>
<td>0</td>
<td>40</td>
</tr>
</tbody>
</table>
Evaluation rules (in brief)

- 4 test conditions (OF, OR, CF, CR) × 3 durations: 12 tracks
- For each test condition: single primary + any number of contrastive systems
- Results in NIST LRE format (text file with one line per trial and 6 fields per line)
- Participants committed to specify whether or not their scores may be interpreted as log-likelihood ratios
- Participants committed to send descriptions of their systems and present them at the Albayzin 2008 LRE workshop
- Systems ranked in each track according to $C_{avg}$
- **Award:** system yielding the least $C_{avg}$ in the CR-30 condition
Schedule (as finally executed)

- Evaluation plan released, registration opens (deadline: July 31)
- Train and development data submitted to registered sites, time for system development
- Evaluation data submitted to registered sites, time for processing evaluation data
- System results and description submitted to organization, analysis of the submitted results
- Keyfile released, results notified to participating sites, time for preparing workshop presentations
- Albayzin 2008 LRE Workshop

May 5 - July 7

Sept 29 - Oct 19 - Nov 12-14

July 7 - Oct 29

Database production
- April-June 2008
- September 2008 (for additional evaluation data)
Results

**Participation:** 4 teams, 13 systems

Two teams (T1: 6 systems and T2: 4 systems) applying state-of-the-art language recognition technology

Average performance in the four test conditions (OF, OR, CF and CR) on the subset of 30-second segments

<table>
<thead>
<tr>
<th></th>
<th>OF-30</th>
<th>OR-30</th>
<th>CF-30</th>
<th>CR-30</th>
</tr>
</thead>
<tbody>
<tr>
<td>pri</td>
<td>0.0946</td>
<td>0.1313</td>
<td>0.0552</td>
<td>0.0778</td>
</tr>
<tr>
<td>con</td>
<td>0.1110</td>
<td></td>
<td></td>
<td>0.0656</td>
</tr>
<tr>
<td>pri</td>
<td></td>
<td>0.2787</td>
<td>0.0556</td>
<td>0.2420</td>
</tr>
<tr>
<td>con</td>
<td>0.2597</td>
<td></td>
<td></td>
<td>0.5389</td>
</tr>
<tr>
<td>pri</td>
<td>0.5035</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results: CR-30 (mandatory condition)

- Best primary (award winner): T1, $C_{avg} = 0.0778$
- Best of all: T1-contrastive, $C_{avg} = 0.0656$
Results: CF-30

- Best performance in CF-30: $C_{avg} = 0.0552$, meaning around 5% EER
- 5.45% EER obtained in independent experiments carried out with our own state-of-the-art system. The same system yielded below 3% EER in the general language recognition task defined in NIST 2007 LRE.
- Performance worse for this task than for the general task defined in NIST 2007 LRE
- Possible issues...
  - Not the same task, not the same data (are results comparable?)
  - Statistical significance (few errors, not many trials)
- ...and possible explanations:
  - Acoustic variability (speakers, channel, background noise)?
  - Phonetic and lexical similarity among target languages?
- In any case, the task seems to be challenging enough to allow further research in language recognition technology
Results: OF-30

- Best performance in OF-30: $C_{avg} = 0.0946$, meaning around 9% EER
- Almost two times the EER in CF-30: impostor trials corresponding to unknown languages introduce a sizeable number of false alarms
- Some unknown languages are being confused with target languages, maybe Portuguese and French?
Results: performance per target language

DET curves for target languages (best systems in CF-30 and OF-30)
Results: performance per target language

<table>
<thead>
<tr>
<th>Segment</th>
<th>Spanish</th>
<th>Catalan</th>
<th>Basque</th>
<th>Galician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>0.0750</td>
<td>0.0167</td>
<td>0.1250</td>
<td>0.0833</td>
</tr>
<tr>
<td>Catalan</td>
<td>0.0083</td>
<td>0.1167</td>
<td>0.0083</td>
<td>0.0000</td>
</tr>
<tr>
<td>Basque</td>
<td>0.0083</td>
<td>0.0000</td>
<td>0.0083</td>
<td>0.0000</td>
</tr>
<tr>
<td>Galician</td>
<td>0.1167</td>
<td>0.0500</td>
<td>0.0083</td>
<td>0.1000</td>
</tr>
</tbody>
</table>

Note. In the paper, error rates were mistaken as costs. An updated version can be downloaded from http://gtts.ehu.es (go to research and then to publications).

Luis J. Rodríguez-Fuentes et al. The Albayzin 2008 LRE (Odyssey 2010, Brno, CZ)
Results: performance per target language

Error rates: $P_{miss}(i)$ in the diagonal, $P_{fa}(i, j)$ outside the diagonal (best system in OF-30)

<table>
<thead>
<tr>
<th>Segment</th>
<th>Spanish</th>
<th>Catalan</th>
<th>Basque</th>
<th>Galician</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>0.0833</td>
<td>0.0083</td>
<td>0.0667</td>
<td>0.0083</td>
</tr>
<tr>
<td>Catalan</td>
<td>0.0083</td>
<td>0.1750</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Basque</td>
<td>0.0083</td>
<td>0.0000</td>
<td>0.0250</td>
<td>0.0000</td>
</tr>
<tr>
<td>Galician</td>
<td>0.1083</td>
<td>0.0417</td>
<td>0.0000</td>
<td>0.1083</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.0667</td>
<td>0.4333</td>
<td>0.1083</td>
<td>0.1417</td>
</tr>
</tbody>
</table>
Pooled DET curves in the OF-30, OF-10 and OF-3 test conditions (best system)

- as expected, worse performance for shorter segments
- EER in OF-3 (around 20%) two times the EER in OF-30 (around 10%)
- similar results for other systems and conditions
Results: development conditions

- Better performance in free-development conditions
- Performance of T1 (blue) and T2 (red) systems not significantly different in CF-30, but...
- T1 restricted system yields 40% worse $C_{avg}$
- T2 restricted system yields 400% worse $C_{avg}$
Conclusions

- LR Evaluation involving the official languages in Spain (Basque, Catalan, Galician and Spanish), using 16kHz speech signals taken from TV broadcasts
- Best system (applying state-of-the-art technology): around 5% EER
- We think that the defined tasks may support further developments in language recognition technology
- Sensitivity to development restrictions depending on the system: 40% vs. 400% increase in cost (interesting for NIST evaluations?)
- Not the same performance among target languages:
  - **Basque**: high performance and low confusion with unknown languages, maybe due to its different origins
  - **Catalan** (and, at a lower degree, also Galician): high confusion with unknown languages
Current work

ALBAYZIN 2010 Language Recognition Evaluation
Current work

ALBAYZIN 2010 Language Recognition Evaluation

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech.
Current work

**ALBAYZIN 2010 Language Recognition Evaluation**

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

- **Schedule:**
  - **July 15:** registration deadline (development data submitted via courier)
ALBAYZIN 2010 Language Recognition Evaluation

**Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

**Schedule:**
- **July 15:** registration deadline (development data submitted via courier)
- **September 27:** evaluation data released
Current work

ALBAYZIN 2010 Language Recognition Evaluation

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

- **Schedule:**
  - **July 15:** registration deadline (development data submitted via courier)
  - **September 27:** evaluation data released
  - **October 11:** deadline for the submission of system results
ALBAYZIN 2010 Language Recognition Evaluation

**Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech.

**Schedule:**
- **July 15:** registration deadline (development data submitted via courier)
- **September 27:** evaluation data released
- **October 11:** deadline for the submission of system results
- **October 30:** keyfile and results released

Register at [http://fala2010.uvigo.es](http://fala2010.uvigo.es) ...and participate !!!
Current work

ALBAYZIN 2010 Language Recognition Evaluation

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

- **Schedule:**
  - **July 15:** registration deadline (development data submitted via courier)
  - **September 27:** evaluation data released
  - **October 11:** deadline for the submission of system results
  - **October 30:** keyfile and results released
  - **November 10-12:** Workshop at FALA 2010 (Vigo, Spain)

Register at [http://fala2010.uvigo.es](http://fala2010.uvigo.es) and participate!!!
Current work

ALBAYZIN 2010 Language Recognition Evaluation

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

- **Schedule:**
  - July 15: registration deadline (development data submitted via courier)
  - September 27: evaluation data released
  - October 11: deadline for the submission of system results
  - October 30: keyfile and results released
  - November 10-12: Workshop at FALA 2010 (Vigo, Spain)

Current work

ALBAYZIN 2010 Language Recognition Evaluation

- **Database:** KALAKA-2, an extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech

- **Schedule:**
  - **July 15:** registration deadline (development data submitted via courier)
  - **September 27:** evaluation data released
  - **October 11:** deadline for the submission of system results
  - **October 30:** keyfile and results released
  - **November 10-12:** Workshop at FALA 2010 (Vigo, Spain)

Register at [http://fala2010.uvigo.es](http://fala2010.uvigo.es) ...and participate !!!