Language identification on broadcast news

Jean-Luc Rouas

Laboratório de sistemas de língua falada

L2F - INESC-ID
L2F Automatic Speech Recognition (ASR) works well for European Portuguese but fails with other languages (including other Portuguese varieties).

Portuguese broadcast news are multilingual: interviews are subtitled in Portuguese, audio in the original language.

To reduce the error rate, identify the spoken language.

Can be useful if several ASR systems are available for most frequent other languages (e.g. English) ⇒ select the appropriate ASR system.
Approaches to language identification

- Acoustic Modeling: MFCCs or SDC features, GMM or SVM models
- P-PRLM: Parallel Phone Recognition followed by Language Modeling
- Prosody: still marginally used
Actual performances: NIST LRE 2005

- Task: verification of the presence of a target language, 7 target languages, spontaneous telephone speech

- Best system: MIT-LL ⇒ 4.2 % EER on 30s. test segments

- Architecture: Fusion of 6 systems
  - 2 acoustic systems:
    - GMM-SDC: GMM modelling on Shifted Delta Cepstra features
    - SVM-SDC: Support vector machine on Shifted Delta Cepstra features
  - 4 PPRLM systems (using 14 phonetic decoders):
    - PPRLM: phone recognition followed by language modeling
    - PPRLM-lattice: phone recognition followed by language modelling on phone lattices
    - PPRSVM-lattice: phone recognition followed by SVM modeling on phone lattices
    - PPR-BT: phone recognition followed by binary tree
The system

Audio signal → Pre-processing → PRLM System → Fusion → Decision

Pre-processing → Acoustic System → Fusion → Decision

Pre-processing → Prosodic System → Fusion → Decision
Pre-Processing

- Automatic segmentation in speech/non-speech
- Automatic speaker clustering ⇒ one decision per speaker (assume one language per speaker)
- (use information on background – not implemented yet)
Acoustic system

- based on acoustic properties of languages
  (≠ phones for ≠ languages)
- does not require annotated data
- Features: MFCC+Delta (8ms)
- Models: GMM
PRLM system

- based on phonotactic properties of languages ($\neq$ phones & $\neq$ phone sequences for $\neq$ languages)
- needs one (or more) phone recognizer (trained on labeled data)
- quite complex when using several phone recognizers...
PRLM system

- Language 1
  - Phone recognizer
  - Language Model 1
  - Language Model ...
  - Language Model N

- Language 2
  - Phone recognizer
  - Language Model 1
  - Language Model ...
  - Language Model N

- Language 3
  - Phone recognizer
  - Language Model 1
  - Language Model ...
  - Language Model N

Audio signal → Feature extraction

Fusion & decision
Prosodic system
Prosodic system
Prosodic system

# | c | v | c | c | v | c | c | v | v | c

# | entièrement | dévorés | par | les | limaces

## State of the art

### Description of the system

Acoustic system

PRLM system

Prosodic system

Data

Experiments

Experiments (2)

Conclusions & Perspectives
Prosodic system

State of the art
Description of the system
Acoustic system
PRLM system
Prosodic system
Data
Experiments
Experiments (2)
Conclusions & Perspectives
Prosodic system

<table>
<thead>
<tr>
<th>CC#</th>
<th>CV</th>
<th>CCCCCV</th>
<th>CCV</th>
<th>CV</th>
<th>CCCCCV</th>
<th>CV</th>
<th>CCV</th>
<th>CV</th>
<th>CCV</th>
<th>CCCCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>entièrement</td>
<td>dévorés</td>
<td>par</td>
<td>les</td>
<td>limaces</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Prosodic system
Prosodic system

<table>
<thead>
<tr>
<th>CC#</th>
<th>CV</th>
<th>CCCCCV</th>
<th>CV</th>
<th>CV</th>
<th>V</th>
<th>CCCCC</th>
<th>CV</th>
<th>CV</th>
<th>CCV</th>
<th>CCCCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>entièrement</td>
<td>dévorés</td>
<td>par</td>
<td>les</td>
<td>limaces</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Waveform plots and audio examples]
State of the art
Description of the system
Acoustic system
PRLM system
Prosodic system
Data
Experiments
Experiments (2)
Conclusions & Perspectives

Prosodic system

<table>
<thead>
<tr>
<th>CC#</th>
<th>CV</th>
<th>CCCCCV</th>
<th>CCV</th>
<th>CV</th>
<th>CCV</th>
<th>V</th>
<th>CCCCCV</th>
<th>CV</th>
<th>CV</th>
<th>CCV</th>
<th>CCCCCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Prosodic system
Data

- All languages from the COST278 database, except English from HUB4

<table>
<thead>
<tr>
<th>Language</th>
<th>Learning set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgian Dutch</td>
<td>117</td>
<td>37</td>
</tr>
<tr>
<td>Czech</td>
<td>114</td>
<td>63</td>
</tr>
<tr>
<td>Galician</td>
<td>149</td>
<td>75</td>
</tr>
<tr>
<td>Greek</td>
<td>139</td>
<td>58</td>
</tr>
<tr>
<td>Croatian</td>
<td>138</td>
<td>33</td>
</tr>
<tr>
<td>Hungarian</td>
<td>143</td>
<td>28</td>
</tr>
<tr>
<td>Portuguese</td>
<td>168</td>
<td>24</td>
</tr>
<tr>
<td>Sloven</td>
<td>101</td>
<td>60</td>
</tr>
<tr>
<td>Slovak</td>
<td>140</td>
<td>27</td>
</tr>
<tr>
<td>American English</td>
<td>221</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>23h50min</strong></td>
<td><strong>7h15min</strong></td>
</tr>
</tbody>
</table>
Experiments

▶ Test the presence of a foreign speaker
▶ “Verification” framework
▶ 789 speaker segments ⇒ 7890 tests:
  • 789 target tests
  • 7101 non-target tests
Results

<table>
<thead>
<tr>
<th></th>
<th>PRLM</th>
<th>MFCC</th>
<th>SDC</th>
<th>Prosody Short</th>
<th>Prosody Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>13.98</td>
<td>23.52</td>
<td>35.75</td>
<td>28.50</td>
<td>40.43</td>
</tr>
</tbody>
</table>
Fusion

- Weighted addition: best results using only PRLM, GMM-MFCC and Short-term prosody

![Graph showing performance comparison between different systems and the fused system.](image)
Impact of the test segments duration

- automatic speaker clustering ⇒ very short segments
  shortest segment : 0.94 s. !

⇒ Investigate the influence of minimum segments duration
  on the performances

- 3 minimum durations tested:
  - segments >10s : 618 test segments
  - segments >20s : 394 test segments
  - segments >30s : 272 test segments
Impact of the test segments duration

EER:

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>over 10s</th>
<th>over 20s</th>
<th>over 30s</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER</td>
<td>12.44</td>
<td>9.28</td>
<td>6.58</td>
<td>5.84</td>
</tr>
</tbody>
</table>
Experiments:
Performances of the Portuguese language verifier

- All languages, all durations:
  ⇒ 39 Target segments, 750 Non-target segments

- EER = 2.55%
  - False detections: 19 (/750)
    - 1 EN, 9 GA, 5 GR, 3 SI, 1 SK
  - Missed: 1 (/39)
    - Music segment (in English...), 2 s.
Experiments:
Verifying Brazilian Portuguese

- Objective: Verify if Brazilian is recognized as Portuguese
- Data from the POSTPOR database (Broadcast News) - 3 recordings (03/01, 04/01, and 05/01/2007)

⇒ 169 test segments (only in Brazilian)

- Use only the PRLM Portuguese language verifier (=Portuguese or reject)

⇒ 100% recognized as Portuguese
Conclusions

- Performances: 5.8 % EER on segments > 30s
  NIST MIT-LL system: 4.2 % EER on 30s segments

- ≠ databases: telephone speech vs. broadcast news
  - telephone speech is likely to have worse quality than BN
  - but more diversity in acoustic conditions on broadcast news (e.g. segments with lot of background noise)

- different constraints on the homogeneity of the segments:
  - in NIST: one speaker per file, 30 s. files
  - in COST: automatic speaker clustering (errors?), files > 30s.

⇒ We need to test the system on a similar database participation to NIST 2007 Language Recognition Evaluation?
Perspectives & Future Work

How to improve the performances?

- PRLM system: use lattices
- Acoustic system: test $\neq$ features (PLP, rasta), $\neq$ models (SVM, NN)
- Fusion: Fuzzy, NN

⇒ Incorporate the LId system in the speech recognition system to measure the improvement

⇒ Design a system to identify Portuguese varieties