A ROBUST SPEECH COMMAND RECOGNIZER FOR EMBEDDED APPLICATIONS

Alexandre Maciel
Arlindo Veiga
Carla Lopes
Cláudio Neves
Fernando Perdigão
José David Lopes
Luís de Sá
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training:
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Human-Machine Interaction

Human talks...

The machine performs:
Automatic Speech Recognition

Speech -> ASR: Automatic Speech Recognition -> Text
Feature Extraction
- Noise reduction techniques
- Voice Activity Detection (VAD)

Models (previously trained)

Decoder $\Rightarrow$ result (textual information or action and a confidence measure)

Result: Próxima, 95.43%
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Noise Reduction Techniques

Algorithm E-AFE:
- based on an ETSI standard (ETSI ES 202 050);
- performs noise-robust feature extraction;

Algorithm E-VAD:
- based on logarithm energy.
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Model Training: **Speech Database**

- Speaker distribution by gender and acoustical environment
- 184 hours of recordings
- 232,000 audio files

<table>
<thead>
<tr>
<th>Gender</th>
<th>TVFL (clean)</th>
<th>TVF (factory)</th>
<th>TVV (vehicle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>103</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>197</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>300</strong></td>
<td><strong>36</strong></td>
<td><strong>32</strong></td>
</tr>
</tbody>
</table>

- DB splitting:
  - 75% for training
  - 20% for testing
  - 5% for development
Model Training: Acoustic Modelling

- Acoustic models based on Hidden Markov Models
- Example of a model for the word “ten” (t E n)

- Number of states
- Transition probabilities
- Probability density function
Model Training: Acoustic Modelling

- Whole-word models
- Monophone models
- Triphone Models

abril  sp 6 b r i l~ sp
acrescentar  sp 6 k r @ S s e~ t a r sp
portagem  sp p u r t a Z e~ j~ sp
próximo  sp p r O s i m u sp

abril  sp 6+b 6-b+r  b-r+i  r-i+l~  i-l~  sp
acrescentar  sp 6+k 6+k-r  k-r+S ... t-a+r a-r sp
portagem  sp p+u p-u+r ... Z-e~+j~  e~j~ sp
próximo  sp p+r  p-r+O ... i-m+u  m-u sp
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training:
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Decoder Step

- 1: Built network;
- 2: Wait for VAD notification;
- 3: Process all audio features;
- 4: Select winner path;
- 5: Test if OOV;
  - If OOV, notify “False Recognition” to API;
  - Else, estimate result probability and send to API
- 6: Go to 2:
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API

Results

Demo

Conclusions
Results

How to measure performance?

Original test file

```
"*/TVFL_FLI_LIM_F46_024@8.lab"

sil
opcoes
sil

"*/TVFL_FLI_LIM_MG6_050@8.lab"

sil
frente
sil

"*/TVFL_FLI_LIM_MH8_219@8.lab"

sil
vai
sil
```

Recognized file

```
"*/TVFL_FLI_LIM_F46_024@8.rec"

sil
doce
sil

dele
doce
sil

"*/TVFL_FLI_LIM_MG6_050@8.rec"

sil
frente
sil

"*/TVFL_FLI_LIM_MH8_219@8.rec"

sil
doze
sil
```
Results

Front-end evaluation:
- ETSI Front-end: 94.88 %
- Efficient Front-End: 96.88%

Acoustic modelling evaluation
- Whole-word models: 96.76%
- Monophone models: 89.28%
- Triphone models: 97.03%
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Demo

Voice controlling presentation

Recognizer working in real-time in a vehicle environment
Summary

- Human-Machine Interaction
- Noise Reduction Techniques
- Model Training:
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions
Conclusions

- The recognizer works in real-time over low performant hardware
- Triphone models are more likely to be used
- The ETSI noise reduction front-end might be biased to the database used in development
Questions?

Thank You

End