A ROBUST SPEECH COMMAND RECOGNIZER FOR EMBEDDED APPLICATIONS

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

INSTITUIÇÕES ASSOCIADAS:

instituto de telecomunicações

creating and sharing knowledge for telecommunications

Summary

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

Summary

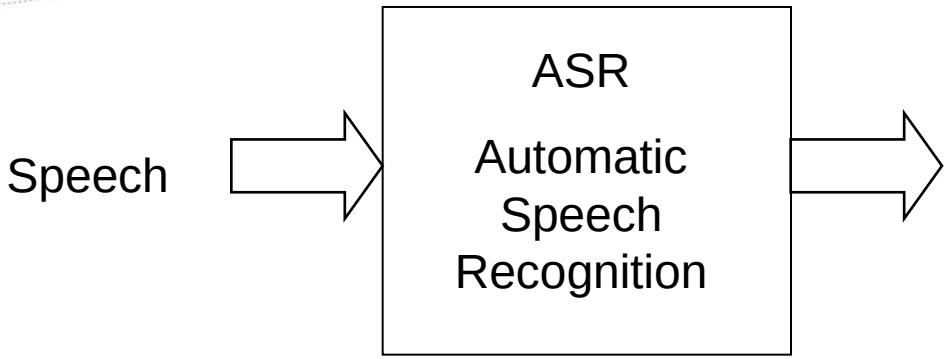
- **Human-Machine Interaction**
- Noise Reduction Techniques
- Model Training
  - Audio Database
  - Acoustic Modelling
- Decoder
- API
- Results
- Demo
- Conclusions

Human-Machine Interaction

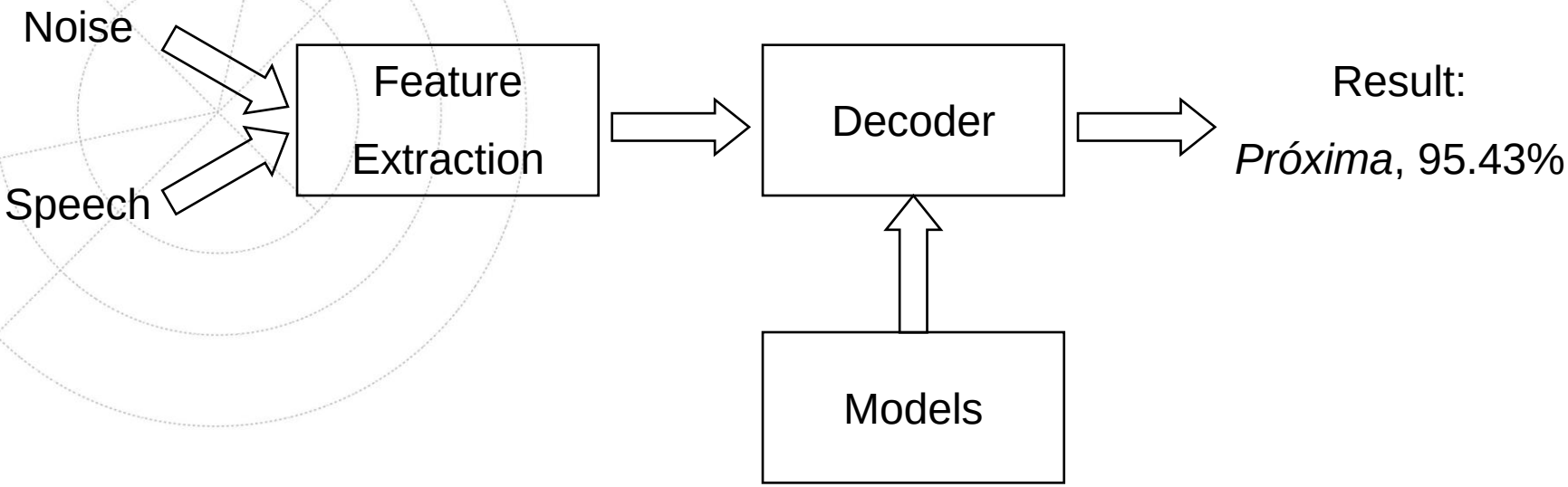
Human talks...

The machine performs:

Automatic Speech Recognition



Human-Machine Interaction



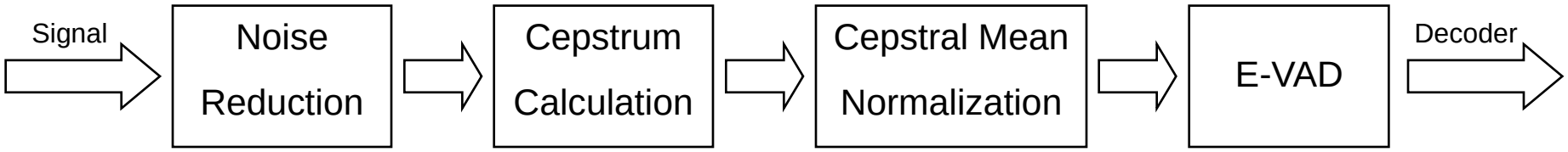
- Feature Extraction
  - Noise reduction techniques
  - Voice Activity Detection (VAD)
- Models (previously trained)
- Decoder ⇒ result (textual information or action and a confidence measure)

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;
  - Neves, C., Veiga, A., Sá, L. and Perdigão, F., 2008. *Efficcient Noise-Robust Speech Recognition Front-End based on the ETSI Standard*. In proc. of ICSP'2008. Beijing, China.
- 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

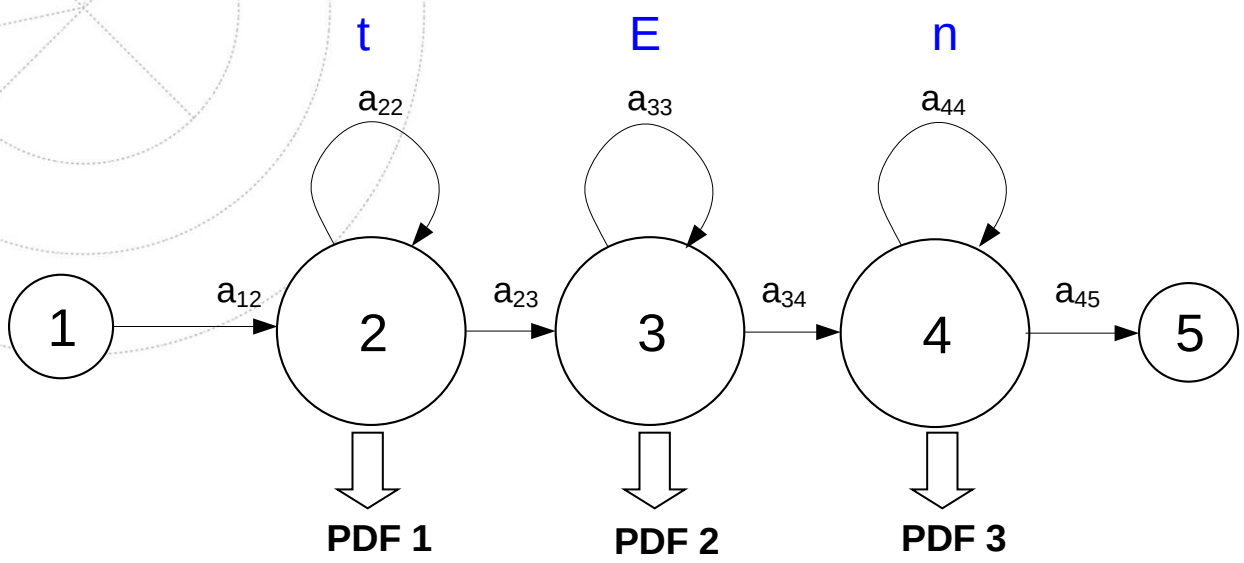
Model Training: Speech Database

- Speaker distribution by gender and acoustical environment
- 184 hours of recordings
- 232,000 audio files
- DB splitting:
  - 75% for traning
  - 20% for testing
  - 5% for development

Gender	*TVFL (clean)*	*TVF (factory)*	*TVV (vehicle)*
Female	103	20	9
Male	197	16	23
Total	300	36	32

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
- Monphone 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

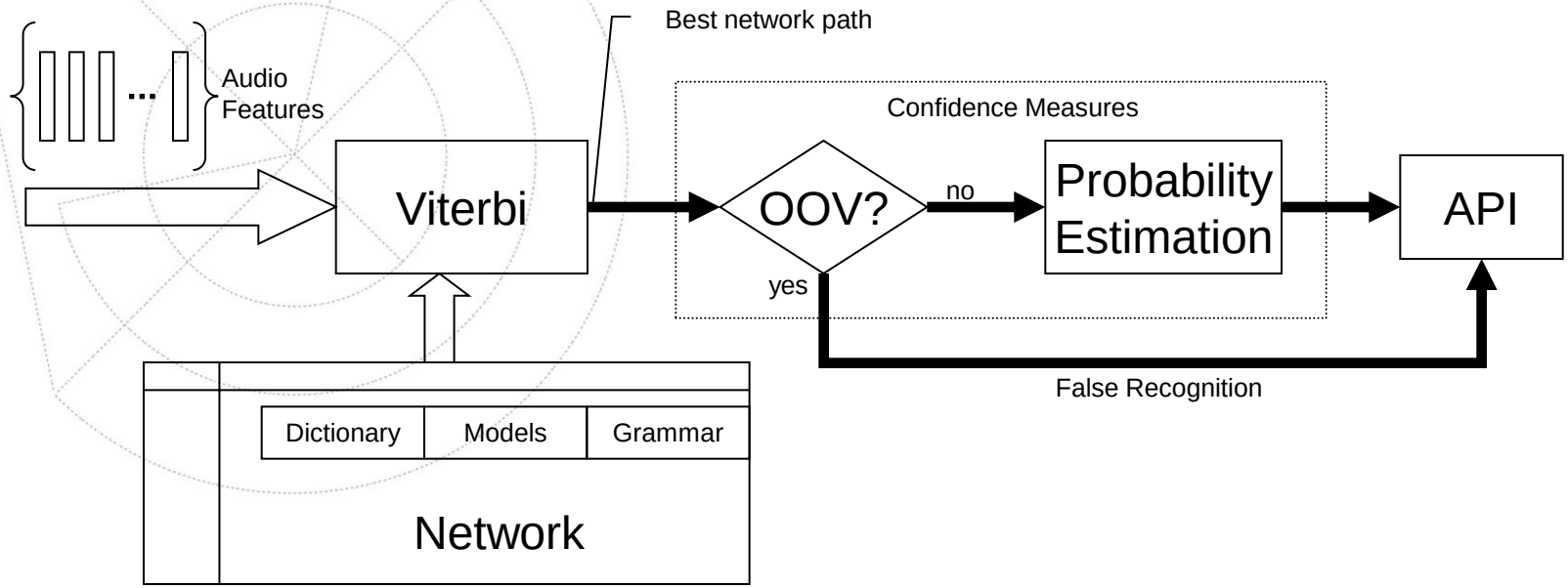
portagem sp p+u p-u+r ... Z-e~+j~ e~-j~ sp

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

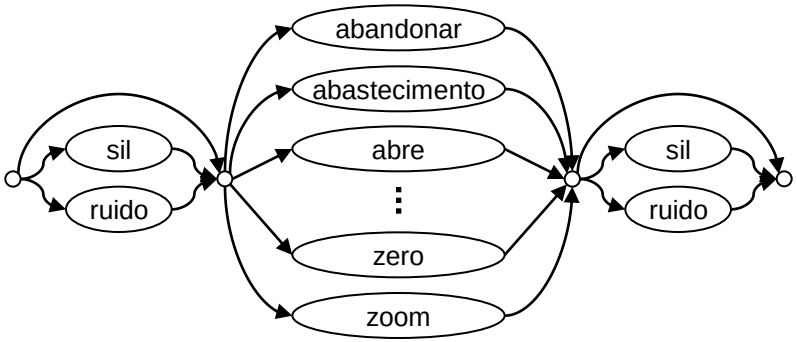
Decoder



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:

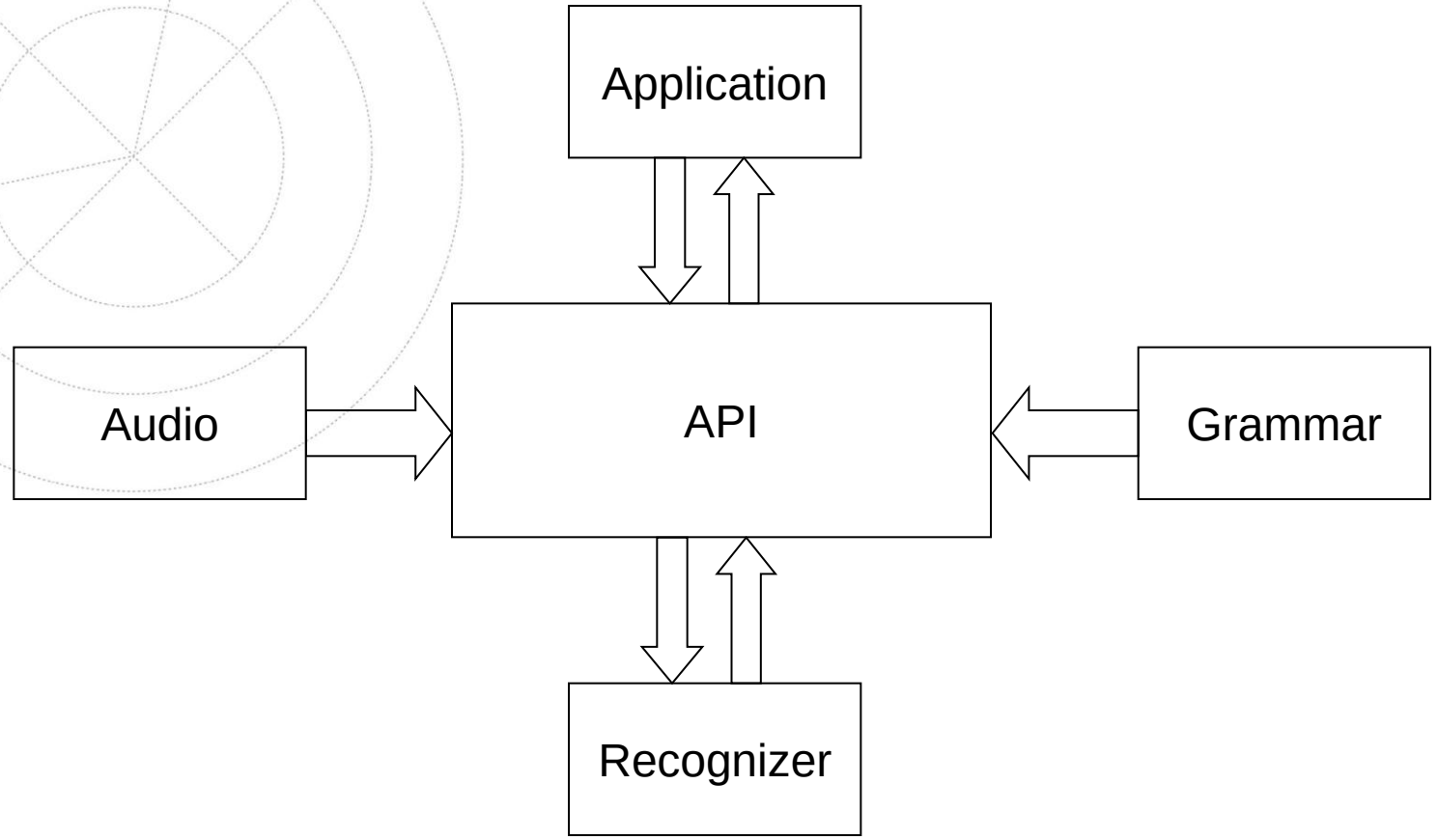
Commands’ Grammar



Summary

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

API



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		Recognized file
“*/TVFL_FLI_LIM_F46_024@8.lab"	sil deletion →	“*/TVFL_FLI_LIM_F46_024@8.rec“
sil		opcoes
opcoes		sil
sil		.
.		
“*/TVFL_FLI_LIM_MG6_050@8.lab"		“*/TVFL_FLI_LIM_MG6_050@8.rec“
sil		sil
frente	Successful recognition	frente
sil		sil
.		.
“*/TVFL_FLI_LIM_MH8_219@8.lab"	Word misrecognized →	“*/TVFL_FLI_LIM_MH8_219@8.rec"
sil		sil
vai		dezoito
sil		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 environement

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 databased used in development

Questions?

:Thank You

End

instituto de telecomunicações

creating and sharing knowledge for telecommunications

INSTITUIÇÕES ASSOCIADAS:

INSTITUTO SUPERIOR TÉCNICO

Faculdade de Ciências e Tecnologia da Universidade de Coimbra

universidade de aveiro

PT Inovação

SIEMENS Communications