EUROPEAN LANGUAGE EQUALITY

META-NET

META-FORUM 2023

Generation of a Large Speech Corpus for Spain Languages using Data Augmentation

José Miguel Herrera (Pangeanic)
j.miguel@pangeanic.com

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu

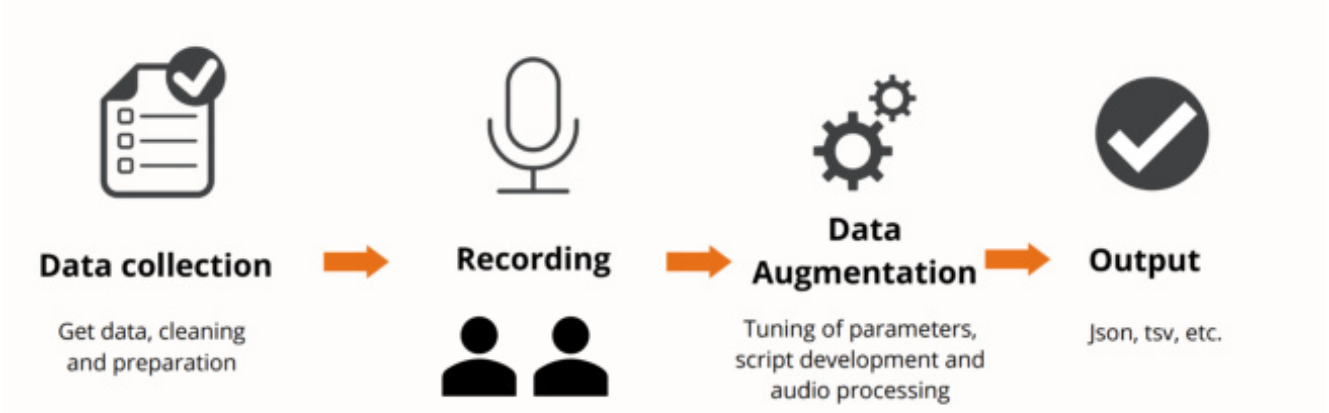
Goal

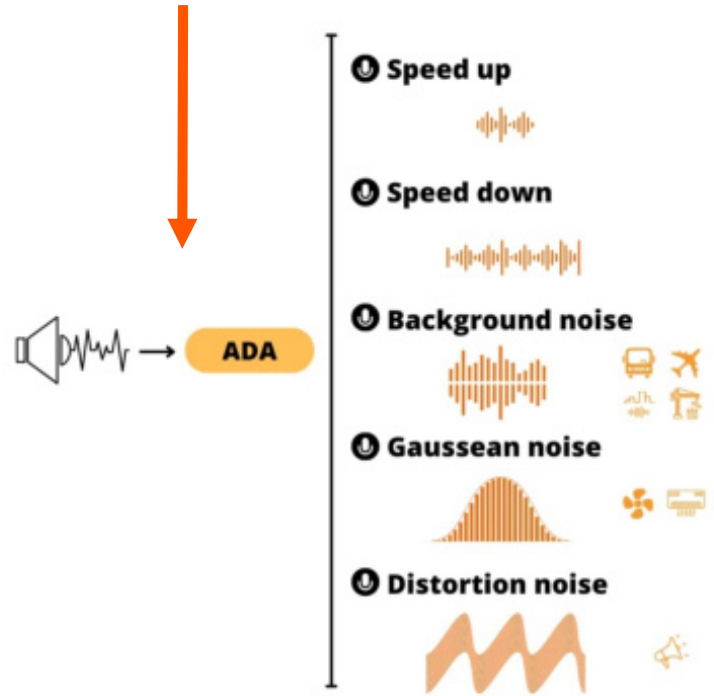
*Generate a **guideline** for building an extensive speech dataset with transcription of Spanish languages through audio data augmentation (ADA) techniques*

Tasks

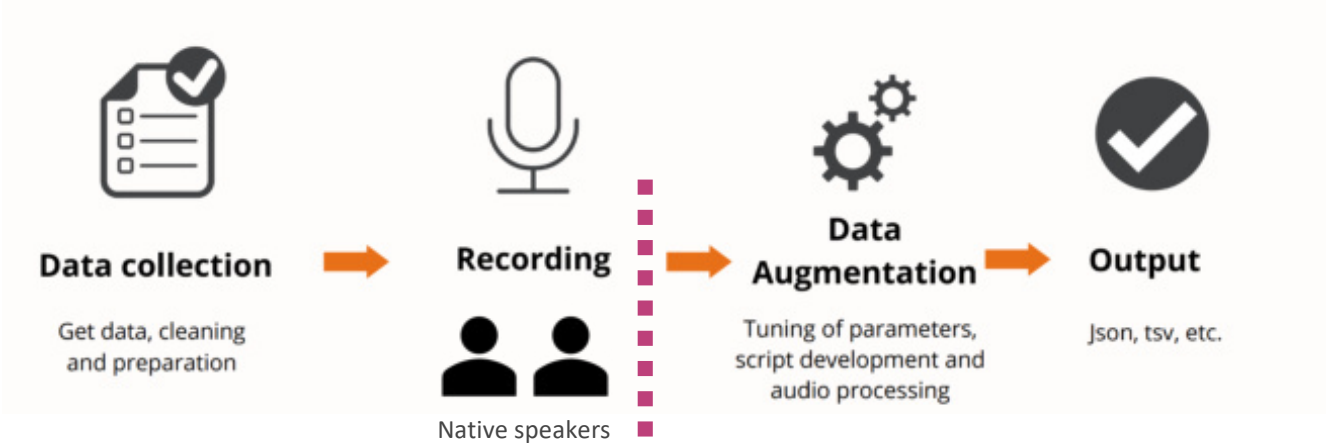
- Generate a guideline
- Create a corpus of four Spanish languages: Galician, Catalan, Asturian and Basque
- Apply the guideline (with ADA) to one of them

Pipeline


Data collection
Get data, cleaning and preparation
Recording
Native speakers
Data Augmentation
Tuning of parameters, script development and audio processing
Output
Json, tsv, etc.


ADA
Speed up
Speed down
Background noise
Gaussean noise
Distortion noise

Pipeline



- We recruited 55 people (32 F - 23 M)
- 182.4 hours
  - Plus associated metadata such as duration, age range and gender.

Language	#audios	hrs	% female/male
Basque	12,231	37.6	60.4% /39.6%
Catalan	20,277	56.7	57.6% /42.2%
Asturian	8,459	25.6	63.4% /36.6%
Galician	22,304	62.5	47.1% /42.9%
Total	**63.271**	**182.4**	

ADA => 20x.

Delivery

- Guideline (PDF report)
  - Experience
  - Challenges
  - Suggestions
  - Analysis of the acceptable range of values for the parameters of each ADA technique

- Datasets (4 Spanish languages), scripts and ADA (Asturian), are publicly available on ELG*

*https://live.european-language-grid.eu

EUROPEAN LANGUAGE EQUALITY

Thank you!

The European Language Equality project has received funding from the European Union under grant agreements № LC-01641480 – 101018166 (ELE) and № LC-01884166 – 101075356 (ELE2).

José Miguel Herrera (Pangeanic)
j.miguel@pangeanic.com

27-06-2023 META-FORUM 2023 – Digital Language Equality for a Multilingual Europe
http://european-language-equality.eu