Speech to Text
Deep Learning agent

Fotis Foukalas and Panagiotis Valatsos

Cogninn

Speech-to-text DL model application

- digiGOV is an Open source pilot project aimed at enhancing the accessibility of public digital services for individuals with disabilities and not only.
- By fine-tuning the **XLSR-Wav2Vec2** model, we strive to improve voice recognition accuracy, ensuring an inclusive user experience that meets high standards.

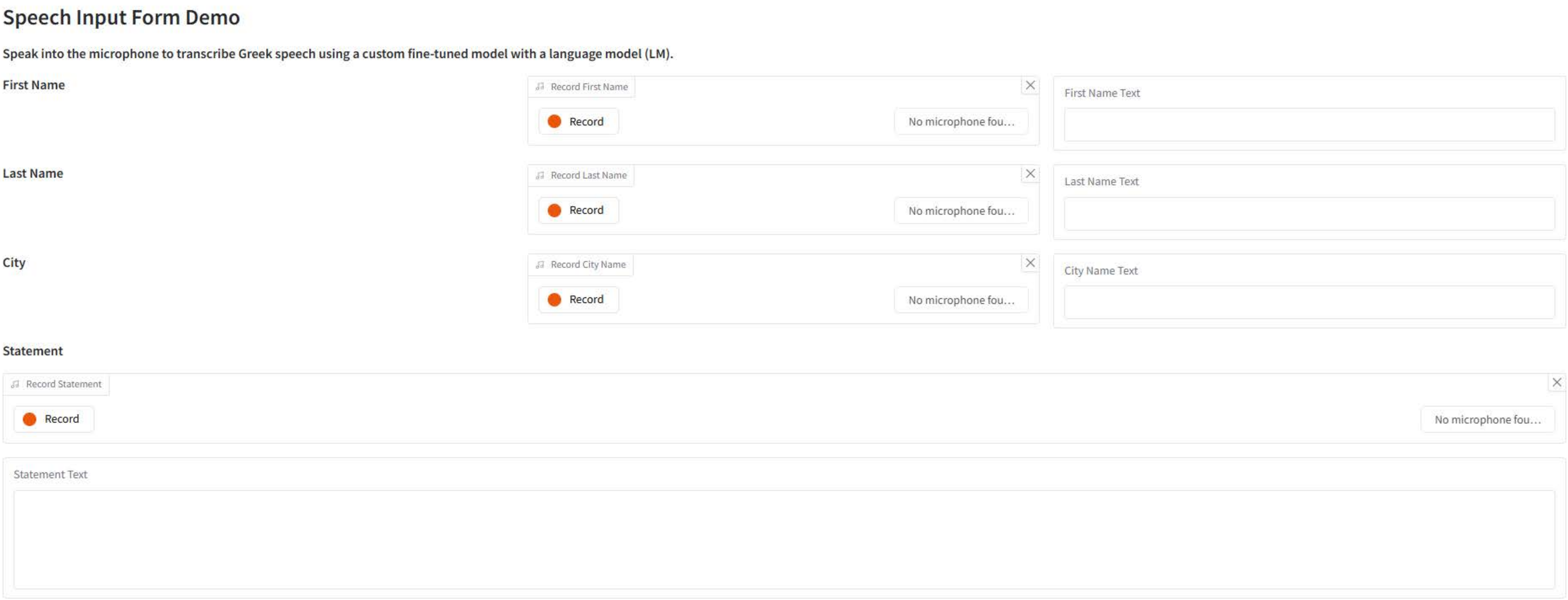
Goals

•**Enhanced Accessibility:** Improve digital service access for individuals with disabilities.

•**User Satisfaction:** Increase usability and satisfaction, particularly for users benefiting from accurate voice recognition.

•**Use:** It can be used to any gov.gr form and any other web applications.

User Interface



Speech Input Form Demo

Speak into the microphone to transcribe Greek speech using a custom fine-tuned model with a language model (LM).

First Name

First Name Text

Last Name

Last Name Text

City

City Name Text

Statement

Statement Text

Validation results

- Common Voice 17, hardest set:
  - Clean test-set: 22% improvement on WER
  - High noise test-set: 15% improvement on WER

Metrics (avg)	Initial checkpoint	Cogninn checkpoint	Average improvement
WER	9.81	7.51	23.44%
CER	2.87	2.98	-0.3%
MER	6.87	5.51	19.79%

Metrics (on CV17)	Initial checkpoint	Cogninn checkpoint	Stressed Initial	Stressed Cogninn
WER	10.02	7.75	48.23	40.79
CER	2.95	3.11	23.11	20.82
MER	7.05	5.68	27.65	23.43

Model

- lighteternal/wav2vec2-large-xlsr-53-greek
  - By the Hellenic Army Academy and the Technical University of Crete
- Trained on :
  - CommonVoice 6.1(EL), 364MB, 2020
  - CSS10 (EL), 121.3MB, 2019
- Fine-tuned further by Cogninn:
  - Improved preprocessing
  - Improved hypothesis creation
  - Added transcription postprocessing

Dataset

- Training Datasets:
  - Common Voice 17 (EL), 720.76MB, 2024
  - Augmented Common Voice 15 (EL), 709.28MB, 2023
  - Augmented Common Voice 19 (EL), 724.35MB, 2024
- Test Datasets:
  - Common Voice 15, 17, 19
- Augmentation:
  - Random noise
  - Change pitch
  - Time stretching
  - Random volume change
  - Vocal Tract Length Perturbation

Fine tuning

- Preprocessing:
  - Improved code consistency between training, evaluation, inference
  - Further trained on newer, wider and larger sets:
    - Common Voice17
    - Augmented Common Voice 15, 19
- Hypothesis creation:
  - Added LM-based processor
    - KenLM-based
  - Improved beam creation function and parameters
    - Search and studies-based
- Postprocessing:
  - Hypothesis re-ranking:
    - Transformer-based
  - Pronunciation guided correction:
    - Custom rules-based
  - Punctuation and Capitalization retrieval:
    - Seq2Seq-based
  - Optimal thresholds
    - Search and studies-based

Demo

- Link:
  - http://62.38.252.170:7800/
  - http://18.192.85.53/

Further improvements:

- Advanced processorwithLM
- Seq2Seq-based Error Correction
- G2P-based Pronunciation Guided Correction
- RAG-based Error Correction

Other ASR services

- Google Cloud Speech-to-Text
  - $0.96/hour
- Amazon Transcribe
  - $0.612/hour
- IBM Watson Speech to Text
  - $0.60/hour
- OpenAI Whisper API
  - $0.36/hour
- Vosk vosk-model-el-gr-0.7:
  - Open-source
  - Accuracy TBD, "not extremely accurate"
  - Older architecture, narrowband

Thank you!

Q&A

Fotis and Panagiotis