Babeș-Bolyai University of Cluj-Napoca
Faculty of Mathematics and Computer Science

Enhancing Romanian Speech Recognition by Using Cross-Lingual Data from Romance Languages

Supervisor
Assist. PhD. Briciu Anamaria

Author
Selegean Victor

Contents:

1. Introduction

1.1 Background

Importance of robust ASR Systems

The gap between systems and technologies available for global languages and languages with modest speaker counts

Limited availability of Romanian Speech data

What will be attempted

1.2 Research Questions

Q1. How does the incorporation of multiple different languages as a basis for Romanian ASR affect the final system's performance?

Q2. If the performance of the ASR systems can be improved, is there a limit to how much Spanish and Italian data we can introduce before the performance starts to degrade?

Q3. If such a limit exists, is there an ideal ratio that maximizes the system's performance?

Q4. How do differing degrees of Italian and Spanish interference in the Romanian ASR systems perform in relation to each other?

1.3 Original Contributions

12 datasets with Romanian data augmented with "mock" Romanian based on Italian and Spanish

12 fine-tuned models for Romanian ASR, based on each dataset

One of the models **deployed** on InferenceEndpoints

Basis for a larger **Romanian ASR corpus**

Android application for interacting with the deployed model

General framework for developing low resource ASR models

2. Theoretical Background

2.1 Theoretical Basis

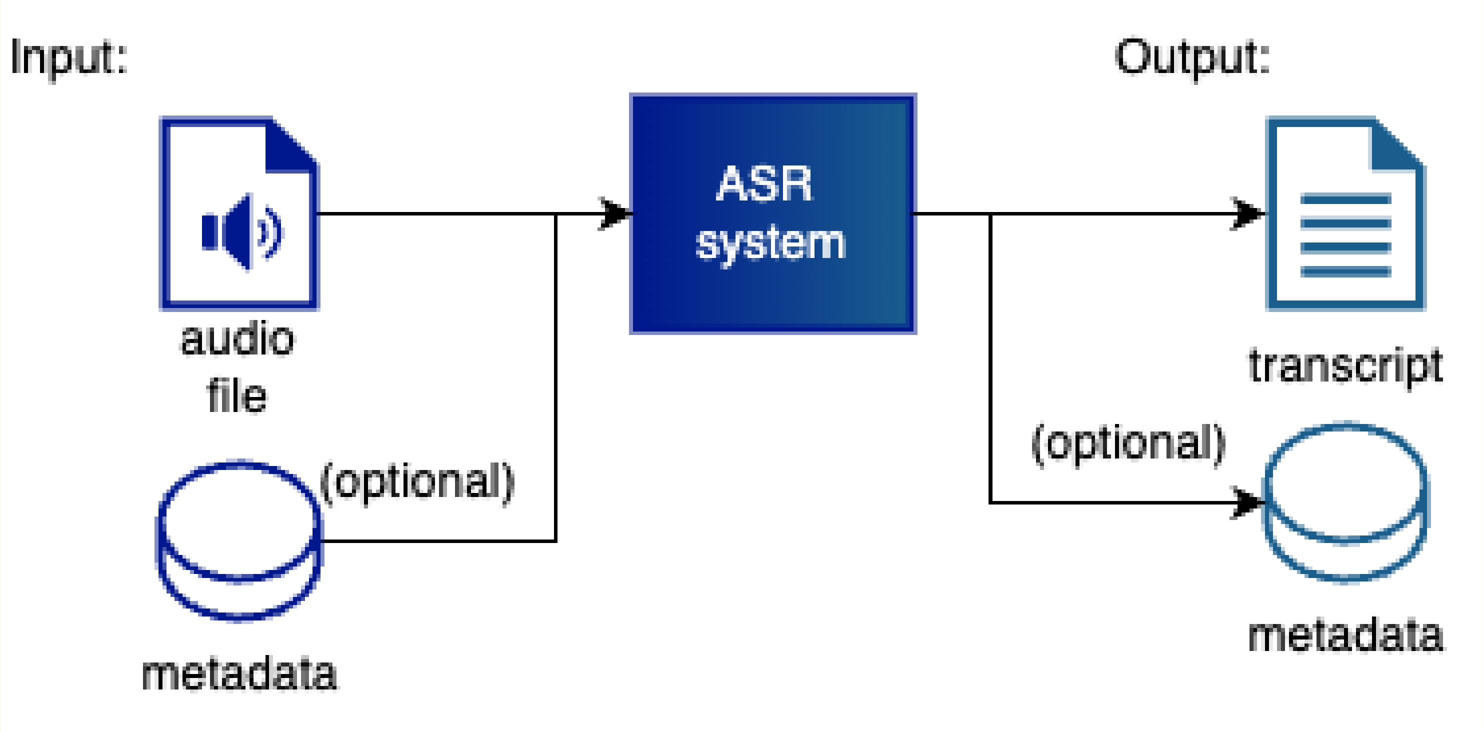


Image created through DrawIO

2.1 Theoretical Basis

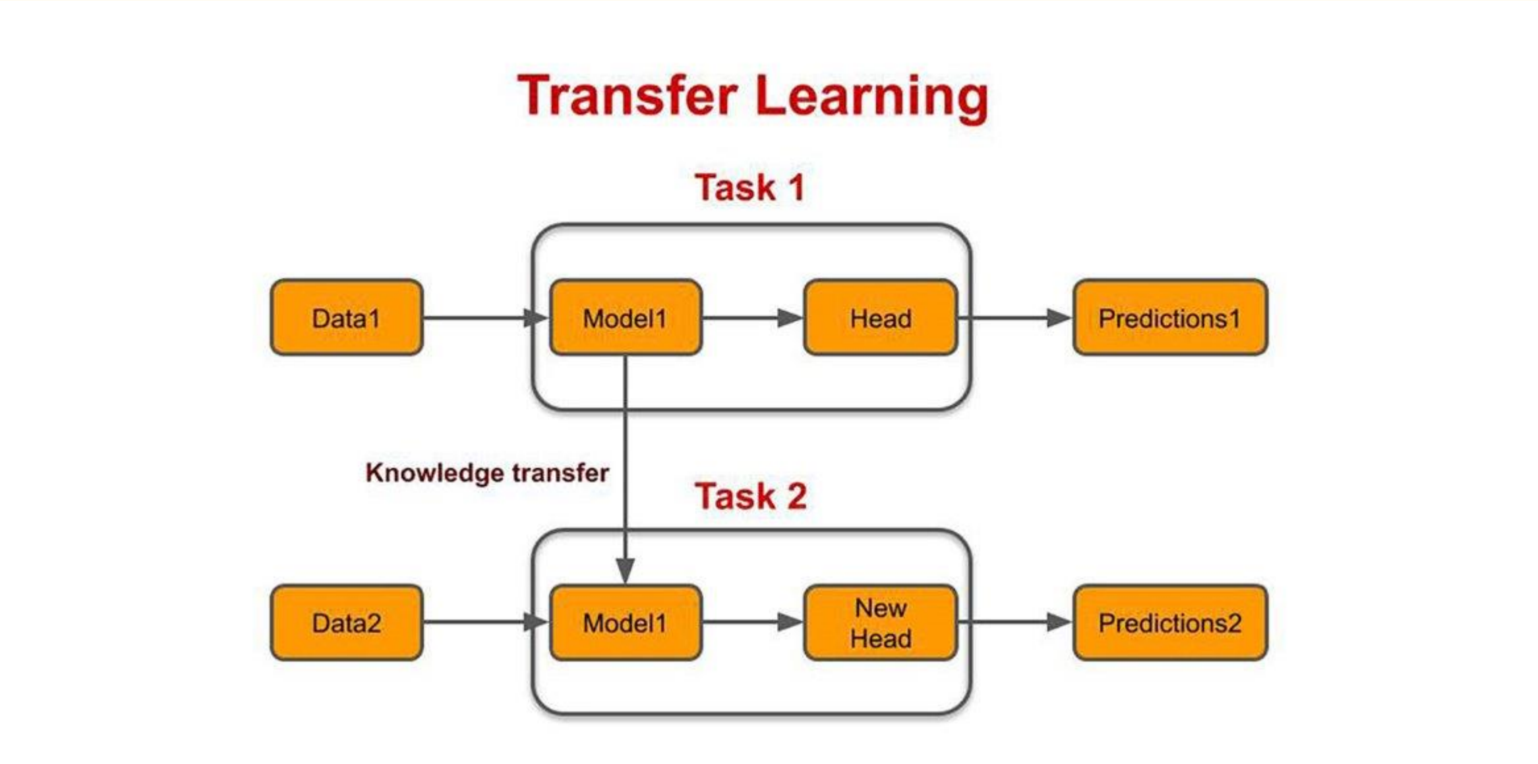


Image from [Hüsein Kaya on Medium]

2.1 Theoretical Basis

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2020)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

Image from International Phonetic Association

English "chart" -> /tʃart/ -> Romanian "ceart"

2.2 Literature Review

2.2.1 Multilingual Representation Learning

- Conneau et al. (2021) - Unsupervised Cross-lingual Representation Learning for Speech Recognition

2.2.2 XLSR between related languages

- Zgank (2019) - Cross-Lingual Speech Recognition Between Languages from the Same Language Family
- Gasan and Păiș (2023) - Investigation of Romanian Speech Recognition Improvement by Incorporating Italian Speech Data

Image from Wikimedia

Image from Official Website

Image from Github

2.2 Literature Review

2.2.3 ASR for Romanian

- Avram, Păiș, and Tufiș (2020) - Towards a Romanian end-to-end automatic speech recognition based on Deepspeech2

Image from Github

2.2.4 ASR Data Sources

- Ardila, Branson, Davis, Henretty, Kohler, Meyer, Morais, Saunders, Tyers, Weber (2020) - Common Voice: A Massively Multilingual Speech Corpus

Image from The Mozilla Blog

3. Automatic Speech Recognition with Cross Linguistic Data

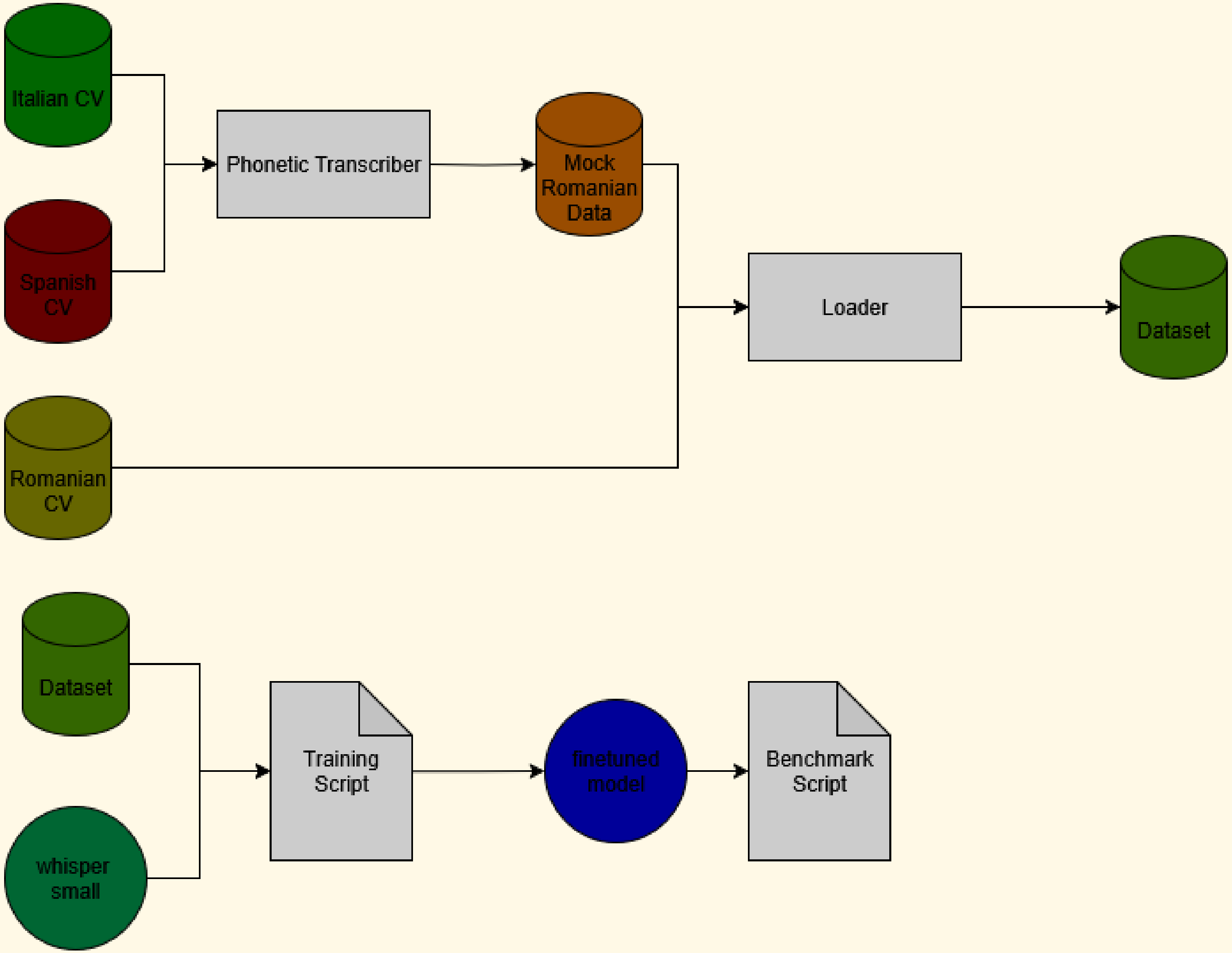


Image created through DrawIO

3.1 Creating a Dataset

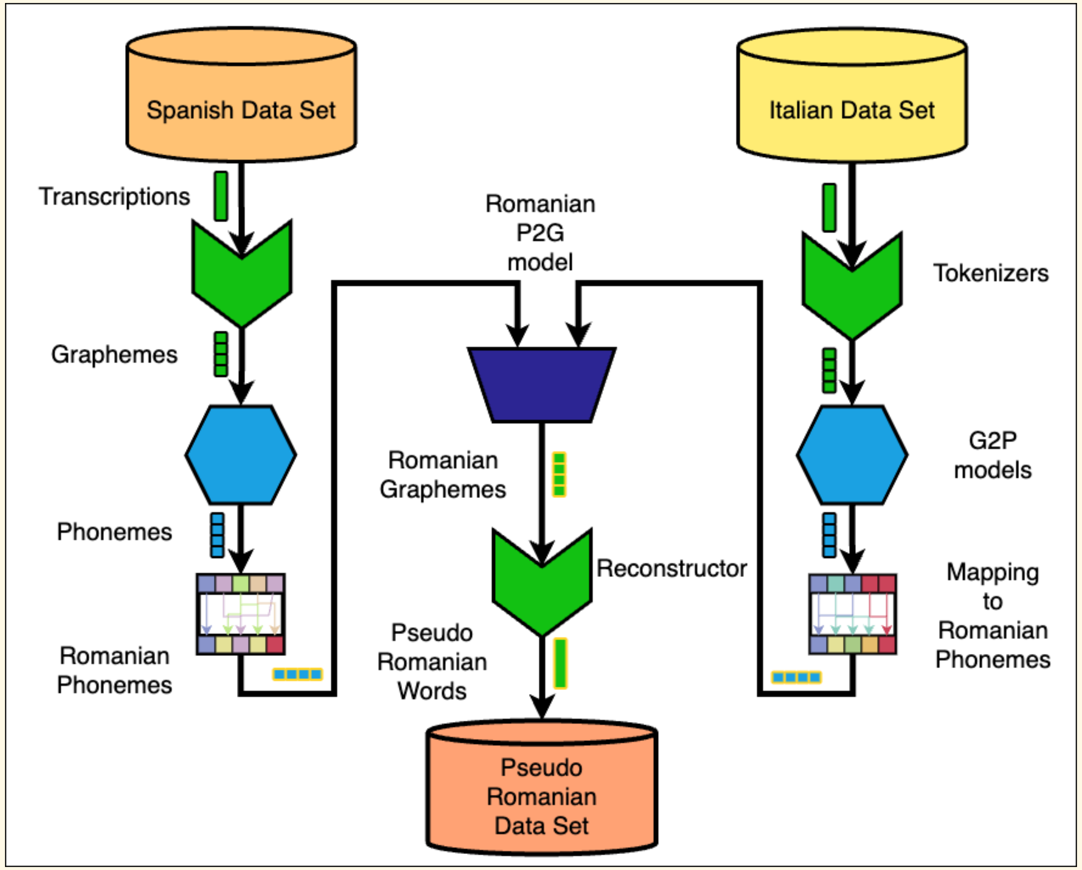


Image created through [DrawIO]

3.1 Creating a Dataset

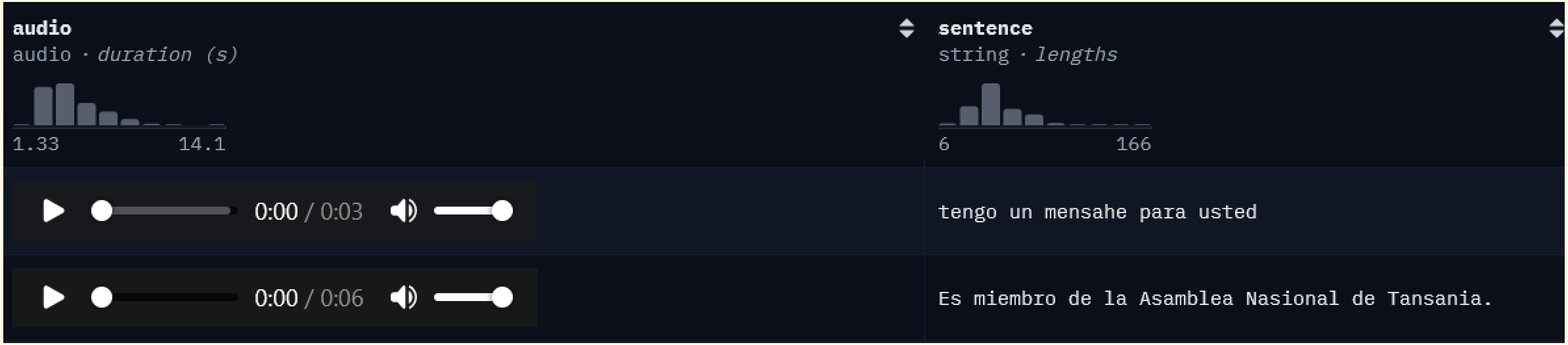
ipa	orthography
string · lengths	string · lengths
1 20	1 19
jew	eu
industrjal	industrial
muʃkjos	muşchios
dojʃpelea̯	doişpelea

Romanian Phonetic Data on [HuggingFace Hub]()

3.1 Creating a Dataset

Converted Italian Data on [Huggingface Hub]

3.1 Creating a Dataset



Converted Spanish Data on [Huggingface Hub]

3.1 Creating a Dataset

Dataset Name	Italian Fraction	Spanish Fraction	Training Set Size
dataset-5k-00it-00sp	0	0	4000
dataset-5k-05it-05sp	5	5	4400
dataset-5k-15it-15sp	15	15	5200
dataset-5k-25it-25sp	25	25	6000
dataset-5k-35it-35sp	35	35	6800
dataset-5k-50it-50sp	50	50	8000
dataset-5k-50it-00sp	50	0	6000
dataset-5k-00it-50sp	0	50	6000
dataset-5k-05it-25sp	5	25	5200
dataset-5k-25it-05sp	25	5	5200
dataset-5k-35it-15sp	35	15	6000
dataset-5k-15it-35sp	15	35	6000

Table created through Latex

3.1 Creating a Dataset

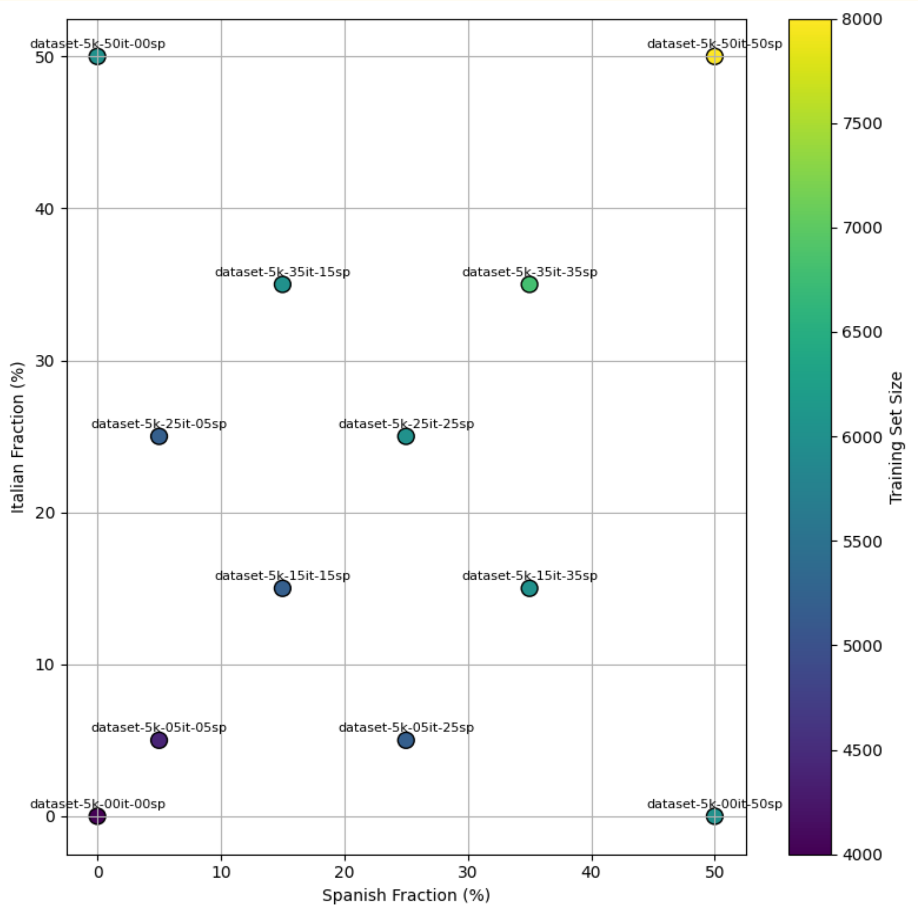


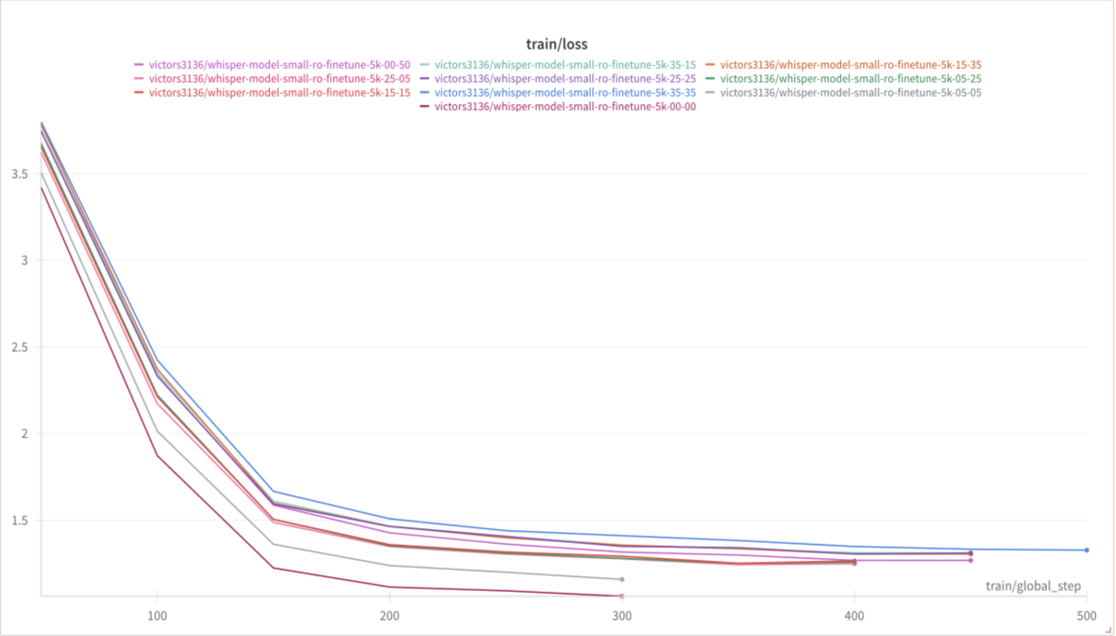
Image created through [Matplotlib]()

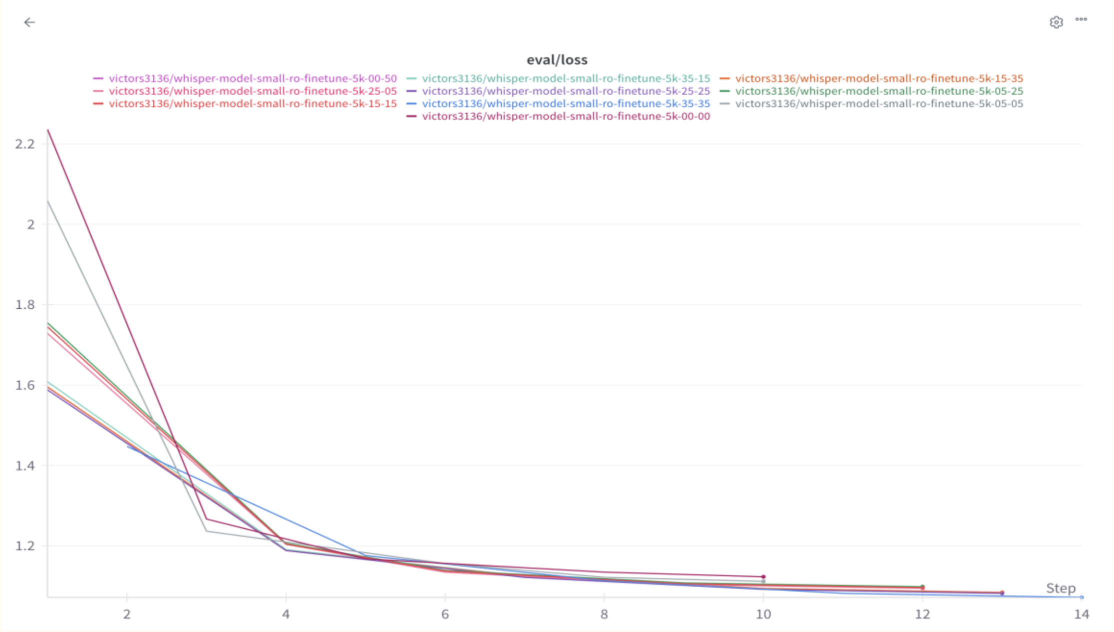
3.2 Training a Model

The Whisper family of models

Image from
[OpenAI]

whisper-small – 244M parameters





Plots generated by [Weights and Biases]

3.3 Comparing Performances

3.3.1 Measuring Results

$$WER(t,m) = \frac{S_W(t,m) + D_W(t,m) + I_W(t,m)}{N_W(t)}$$

where
$t :=$ transcription
$m :=$ model's output
$S_W(t,m) :=$ word substitutions required to turn m into t
$D_W(t,m) :=$ word deletions required to turn m into t
$I_W(t,m) :=$ word insertions required to turn m into t
$N_W(t) :=$ word count of t

$$CER(t,m) = \frac{S_C(t,m) + D_C(t,m) + I_C(t,m)}{N_C(t)}$$

where
$t :=$ transcription
$m :=$ model's output
$S_C(t,m) :=$ character substitutions required to turn m into t
$D_C(t,m) :=$ character deletions required to turn m into t
$I_C(t,m) :=$ character insertions required to turn m into t
$N_C(t) :=$ character count of t

$$RWER_b(t,m) = \frac{WER(t,m)}{WER(t,m_b)}$$

$$RCER_b(t,m) = \frac{CER(t,m)}{CER(t,m_b)}$$

where
$t :=$ transcription
$m :=$ model's output
$b :=$ baseline model
$m_b :=$ baseline model's output

Formulas created using [Latex]

3.3 Comparing Performances

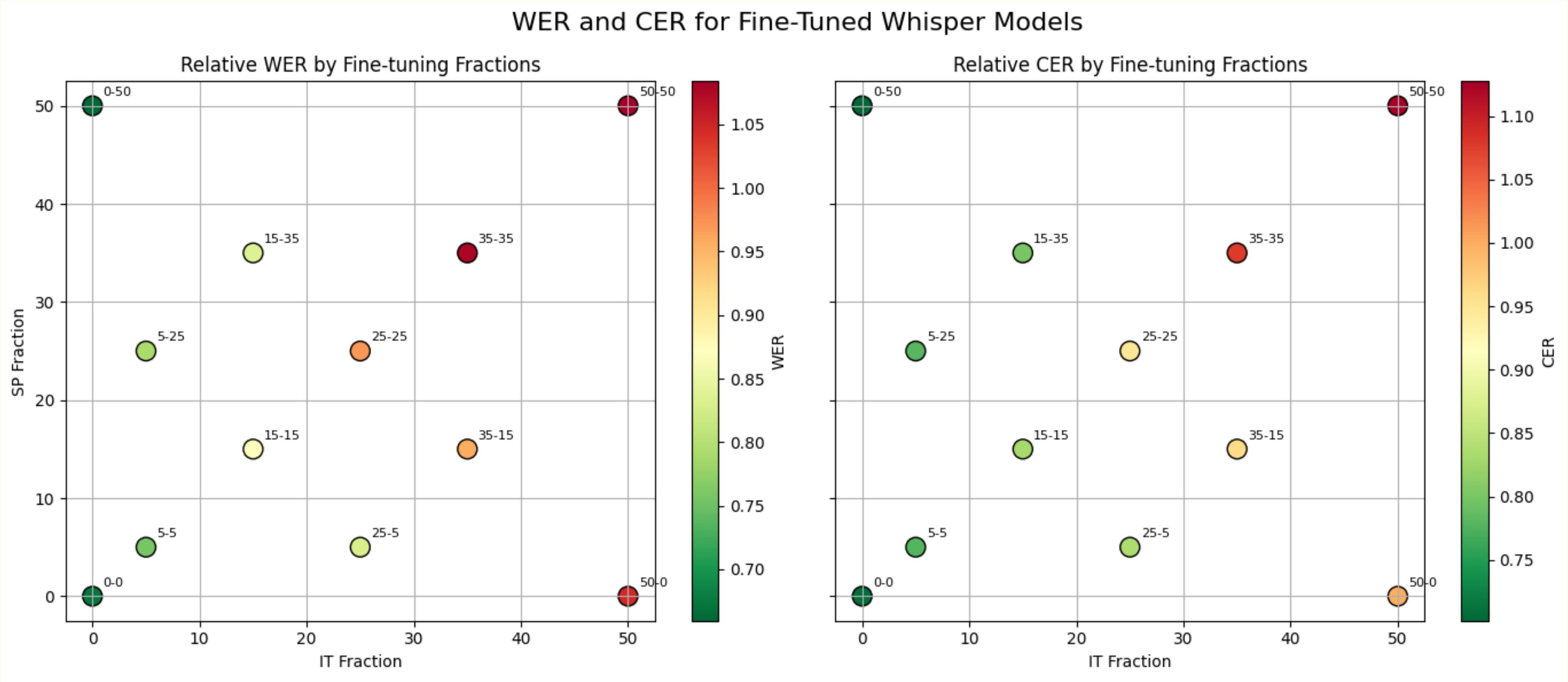
3.3.2 Interpreting Results

Model ID	WER	Relative WER	CER	Relative CER
finetune-5k-00-50	2.96	0.65	2.01	0.75
finetune-5k-00-00	3.31	0.73	2.06	0.77
finetune-5k-05-25	3.39	0.75	2.15	0.80
finetune-5k-05-05	3.47	0.77	2.22	0.83
finetune-5k-15-35	3.88	0.86	2.39	0.90
finetune-5k-15-15	4.11	0.91	2.43	0.91
whisper-small	4.49	1	2.66	1
finetune-5k-35-15	4.79	1.06	2.86	1.07
finetune-5k-25-05	4.79	1.06	2.75	1.03
finetune-5k-25-25	4.93	1.09	2.82	1.06
finetune-5k-50-50	5.35	1.19	3.15	1.18
finetune-5k-35-35	5.58	1.24	3.18	1.19
finetune-5k-50-00	5.89	1.31	3.11	1.17

Table created using Latex

3.3 Comparing Performances

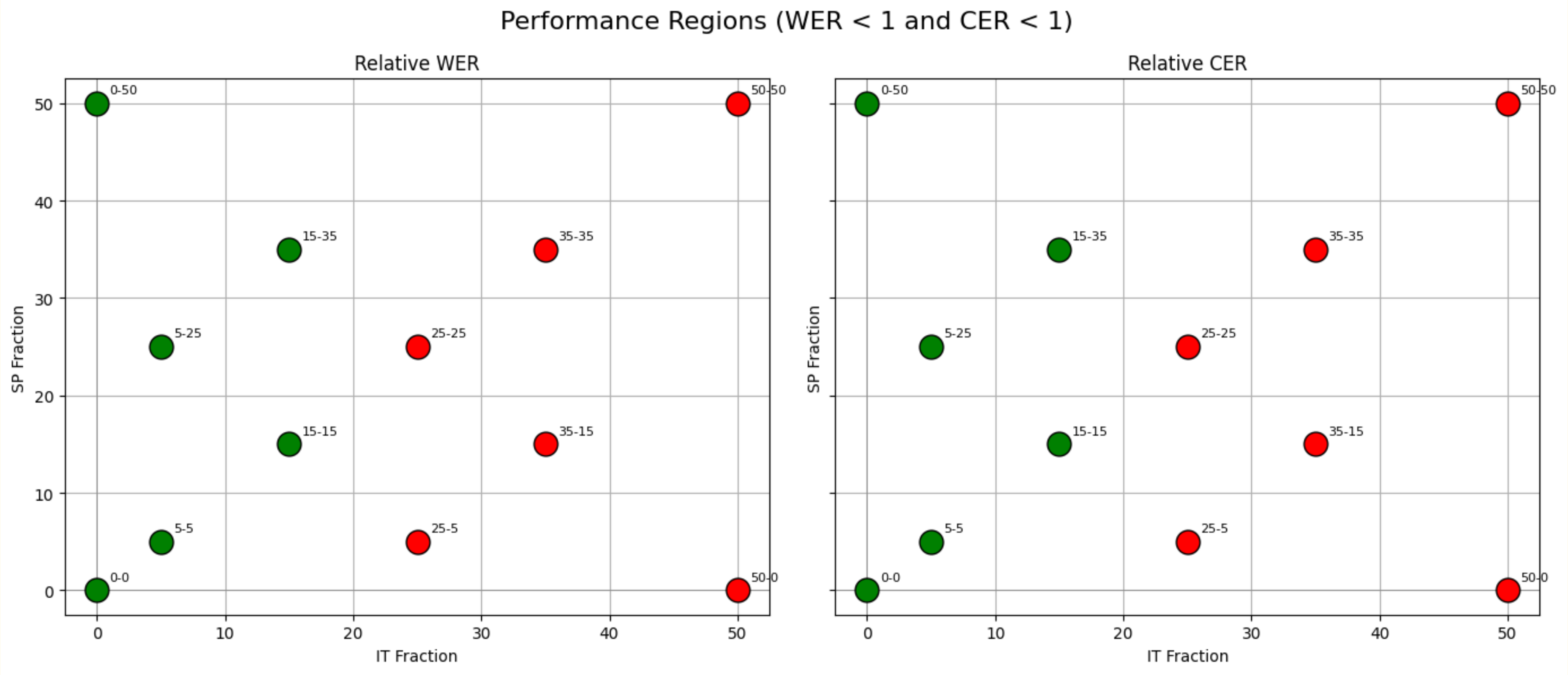
3.3.2 Interpreting Results



Plots created using [Matplotlib]

3.3 Comparing Performances

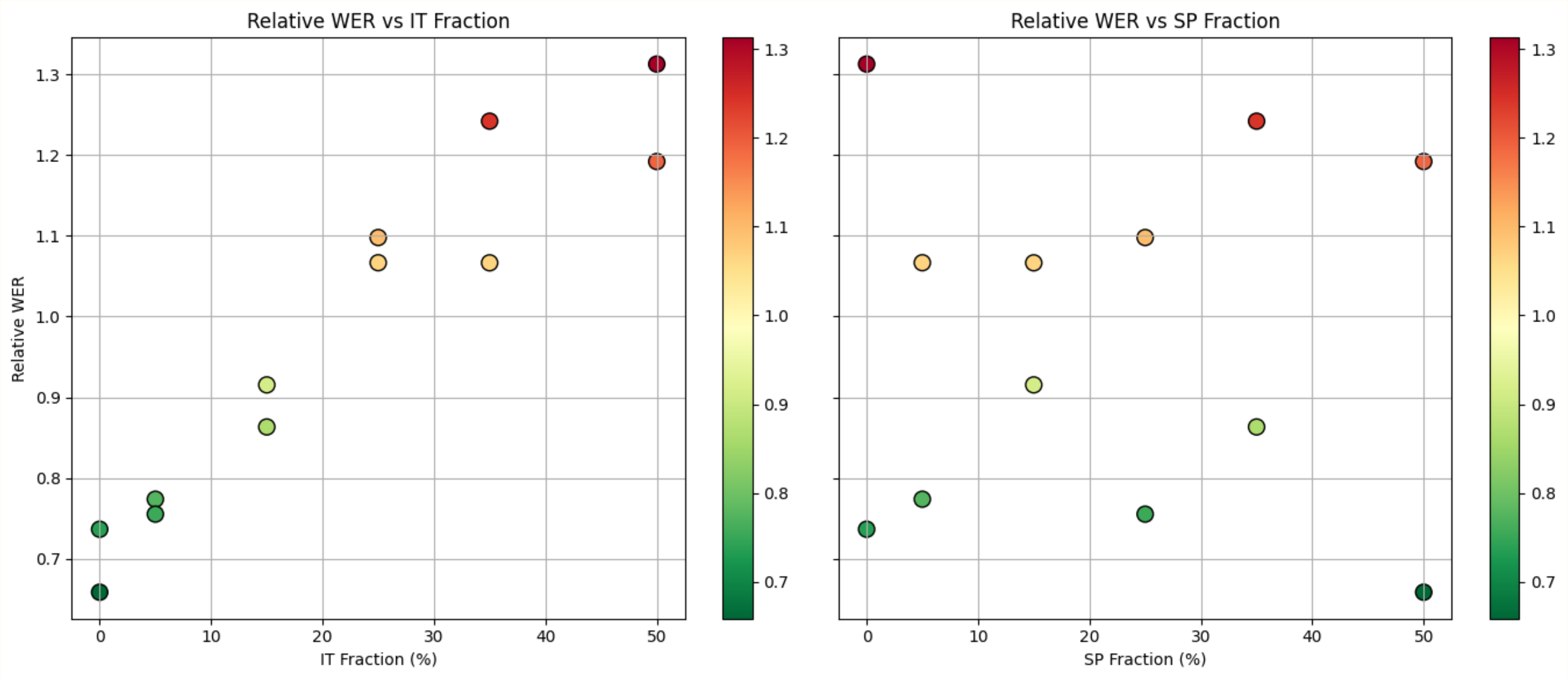
3.3.2 Interpreting Results



Plots created using Matplotlib

3.3 Comparing Performances

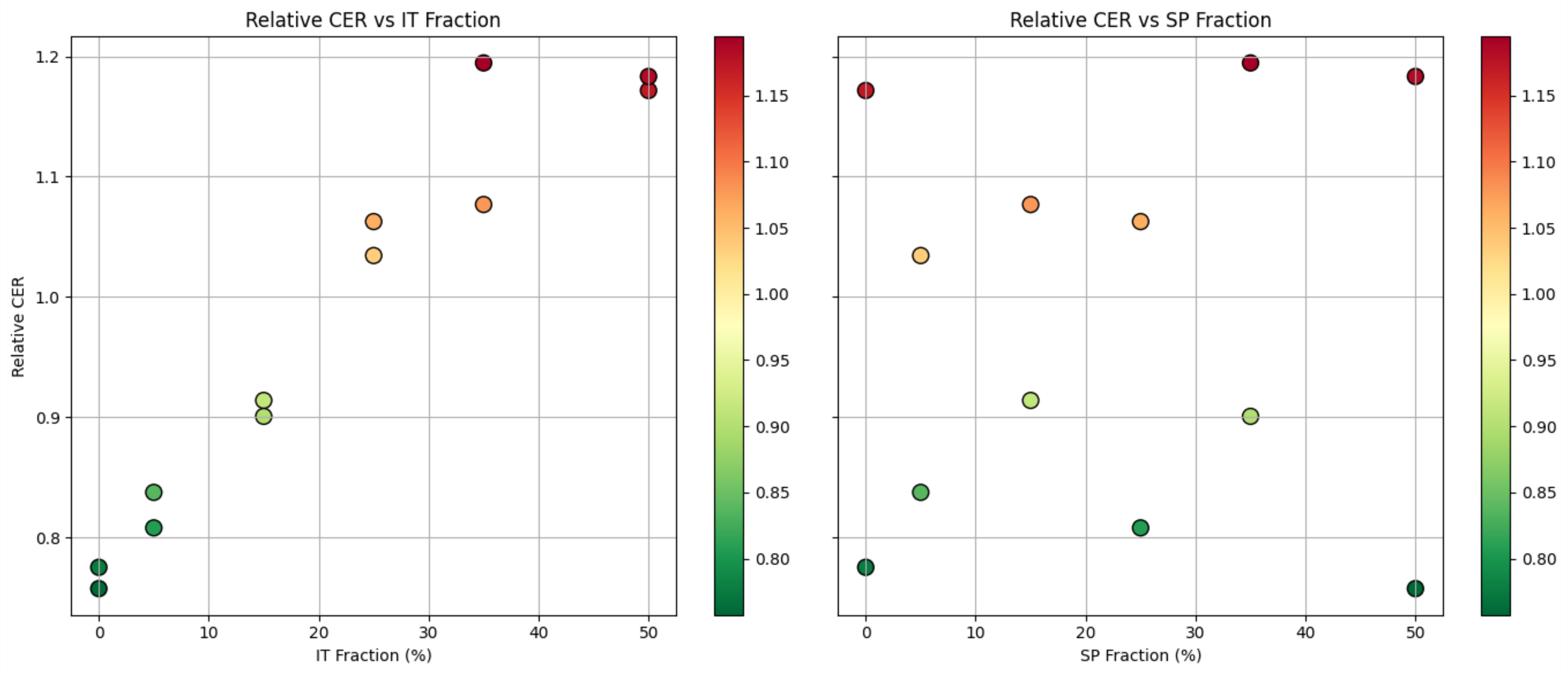
3.3.2 Interpreting Results



Plots created using [Matplotlib]

3.3 Comparing Performances

3.3.2 Interpreting Results



Plots created using [Matplotlib]

4. Application

4.1 Overview

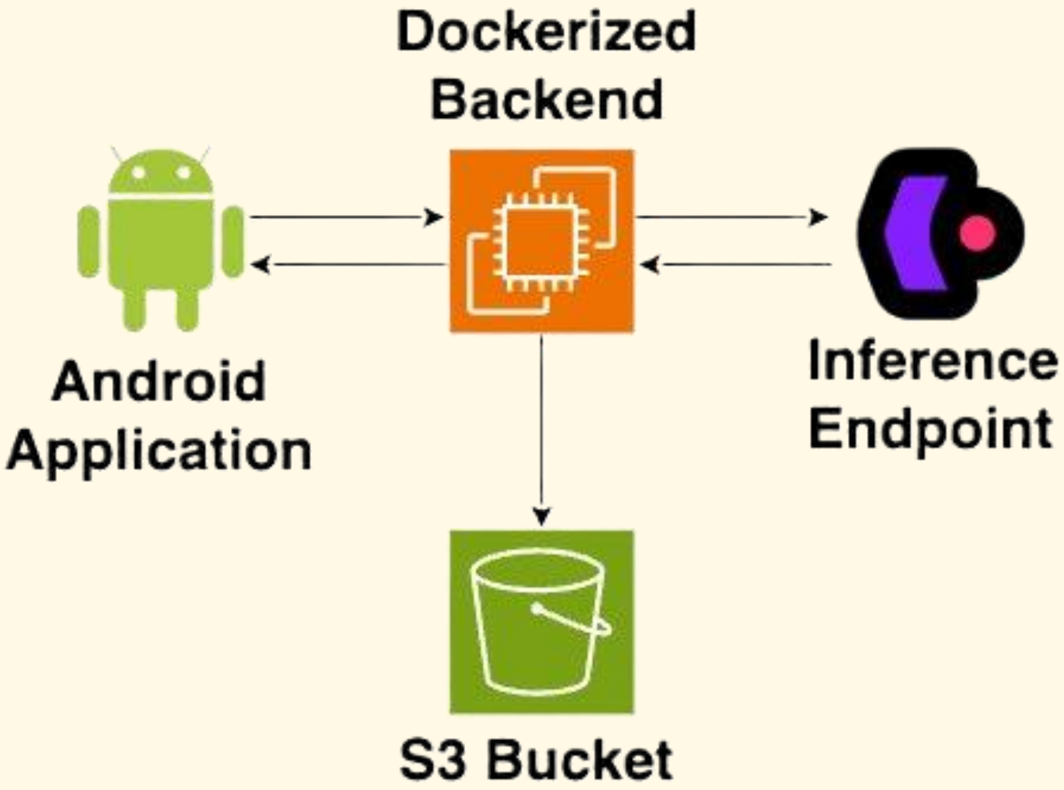


Diagram created through [DrawIO]

4.2 Client

Image from Wikimedia

Image from Wikimedia

Image from Brand PNG Logo

4.2 Client

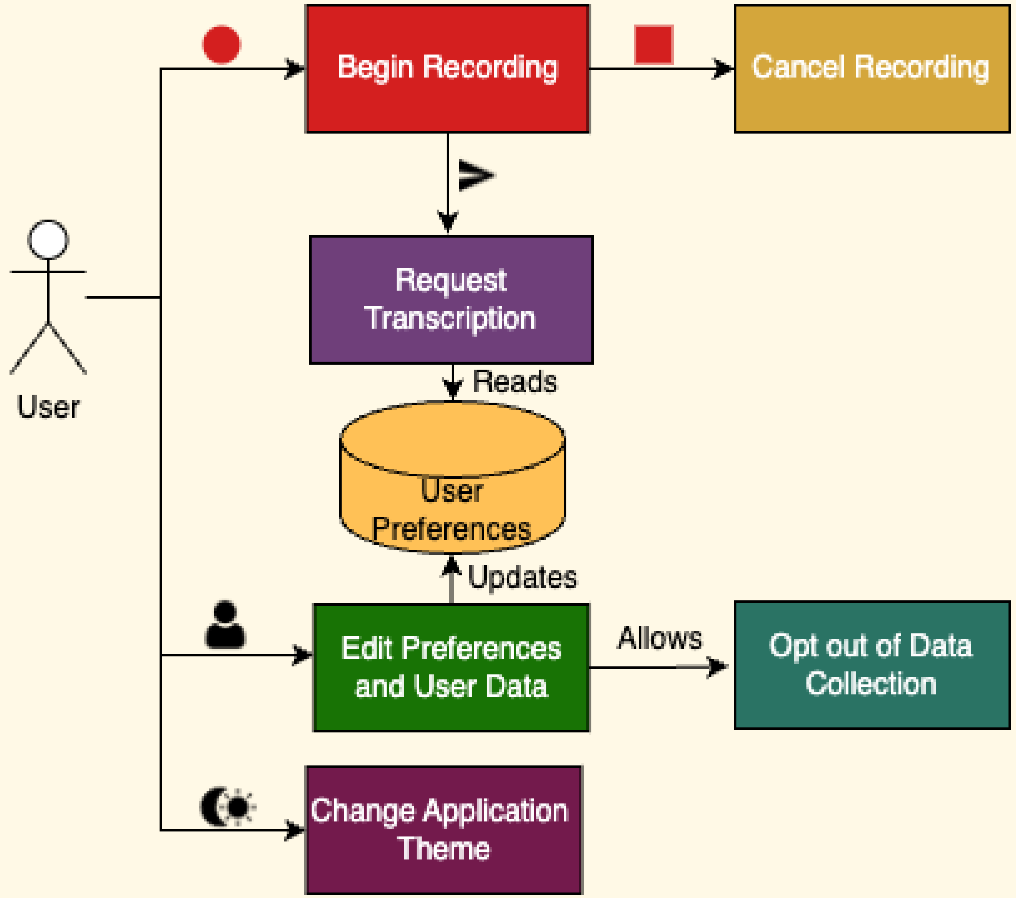
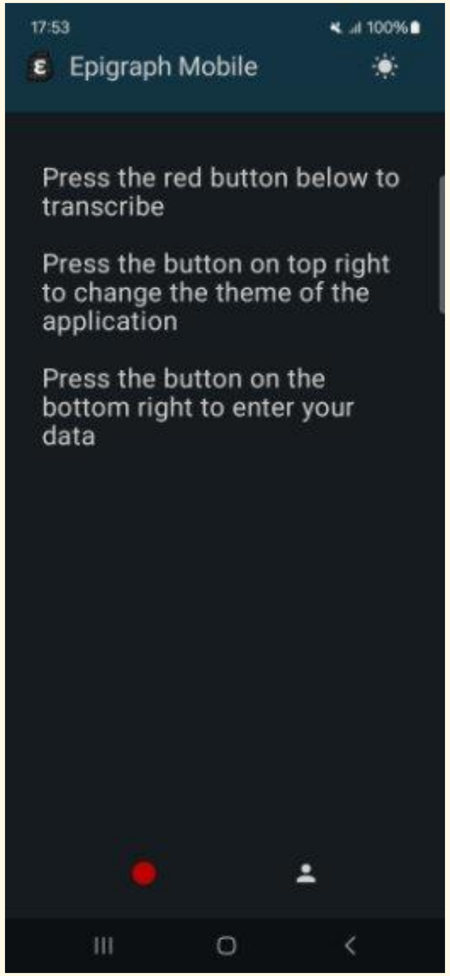


Image created through [DrawIO]



Screenshot of Welcome Screen

4.3 Server

Image from AWS Icons

Image from Official Website

Image from Official Website

4.3 Server

This server powers the Epigraph transcription API. To use it, make a **POST** request to the `/transcribe/` endpoint with your audio data.

The request must be a **multipart/form-data** POST with the following structure:

```
interface TranscriptionRequest {
  file: UploadFile; // .m4a format, max 30 seconds
  age: string?
  gender: "man" | "woman" | "other";
  consent: "true" | "false";
}
```

Download the mobile app here: epigraph-mobile.apk

Landing page of Epigraph Online

4.3 Server

Metadata (3)

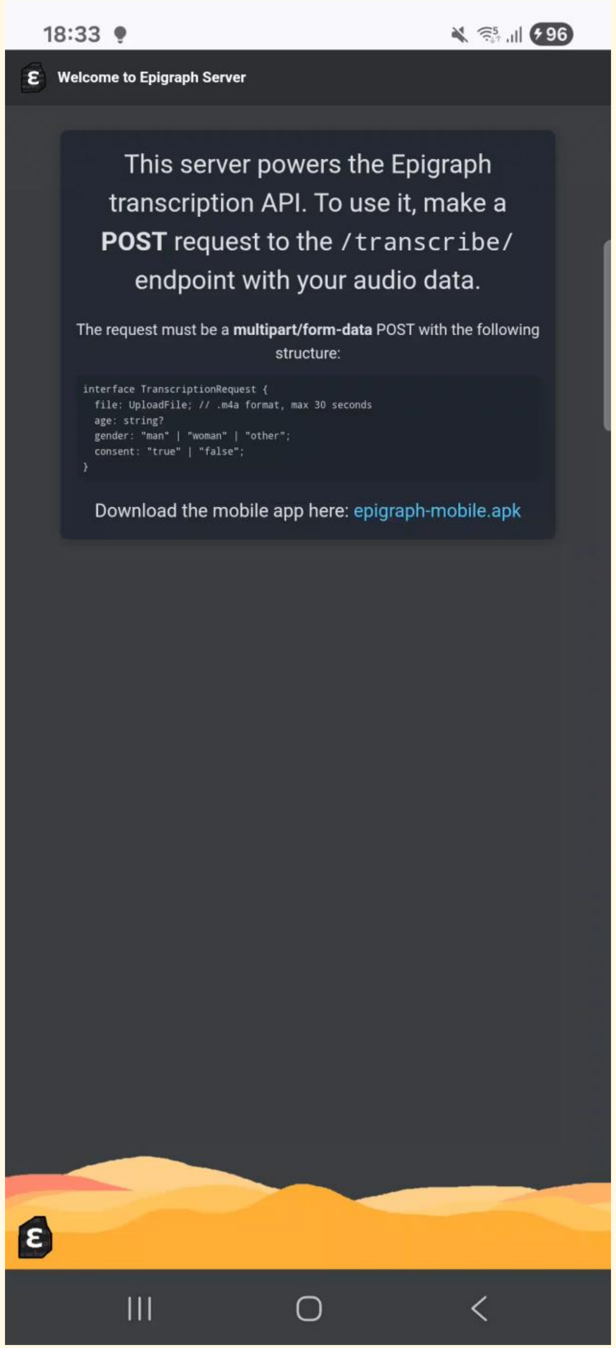
Metadata is optional information provided as a name-value (key-value) pair. Learn more

Type	Key	Value
System defined	Content-Type	audio/m4a
User defined	x-amz-meta-age	80
User defined	x-amz-meta-gender	man

Metadata of an audio file stored on S3

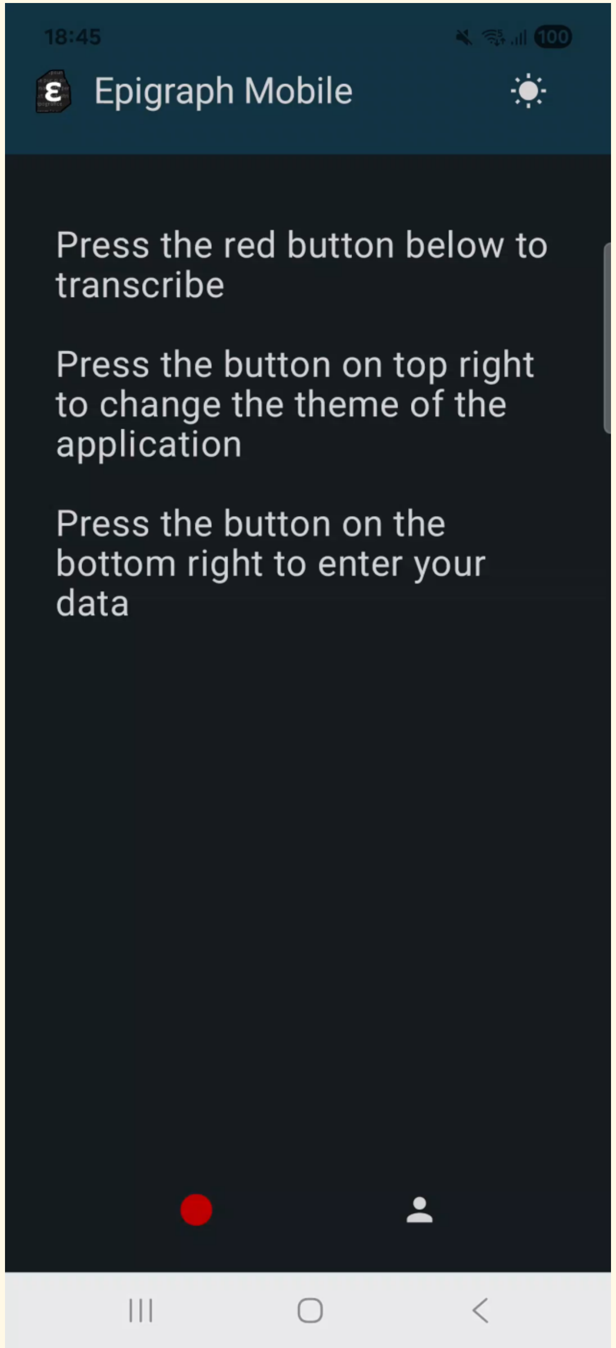
4.4 Demo

4.4.1 Installation. Setup.



4.4 Demo

4.4.2 Transcribing Audio



5. Conclusions

5.1 Answers to Research Questions

Q1. How does the incorporation of multiple different languages as a basis for Romanian ASR affect the final system's performance?

Q2. If the performance of the ASR systems can be improved, is there a limit to how much Spanish and Italian data we can introduce before the performance starts to degrade?

Q3. If such a limit exists, is there an ideal ratio that maximizes the system'sperformance?

Q4. How do differing degrees of Italian and Spanish interference in the Romanian ASR systems perform in relation to each other?

5.2 Further Development

Larger models, longer training times

Targeted design of experiment

Data gathering through the deployed application

Thank you!