AI for Social Good: Bridging Communities through Spoken Language Models

Ayan Basak, Sanket Gupta, Shikhar Kanaskar, Harshit Pant, Jai Trivedi, Matthew A. Lanham
Purdue University, Mitchell E. Daniels, Jr. School of Business
basak0@purdue.edu; gupta795@purdue.edu; skanaska@purdue.edu; hpant@purdue.edu; trivedij@purdue.edu; lanhamm@purdue.edu

BUSINESS PROBLEM FRAMING

- More than 50% of the languages in the world have no digital footprint
- 25% of the world's people are left out because of language-related barriers

How can we bridge this gap?

- SIL works with communities worldwide to develop language solutions that expand their possibilities for a better life
- Partnering with SIL, we have built a robust architecture to train and scale a **language model on AWS**; which takes **audio as an input** and **returns the name of the language being spoken** in the audio

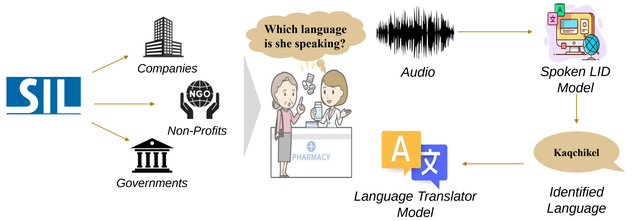


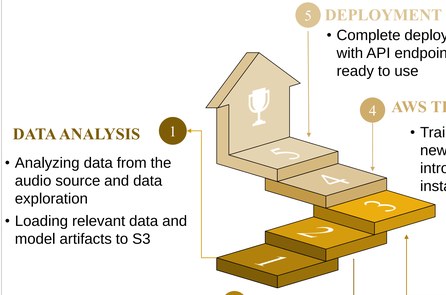
Fig 1. SIL's clients and ways in which they can deploy our solution

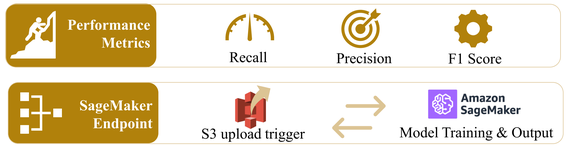
- It is important for any transcription model to first understand the language that the person is trying to communicate in. Our architecture aims to be that rudimentary step in trying to detect the spoken language

- SIL will deploy this model in different countries at locations such as:

- SIL also plans to serve the wider social good community by leveraging the architecture to enable language preservation to aid education efforts for displaced and remote communities
- ***Constraints**: We will be building an operationalized pipeline to model 6 ancient Mayan languages that have negligible digital footprint, and later SIL will scale it for other languages*

ANALYTICS PROBLEM FRAMING





DATA

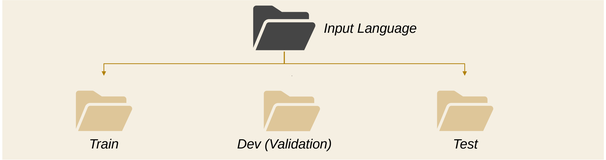


Fig.2. Hugging Face Folder Structure

	Languages	Kaqchikel, Q'eqchi', Q'anjob'al, Mam, K'iche', Spanish
	Training Data	**8,024 files ~ 11.34 hours**
	Validation Data	**299 files ~ 0.45 hours**
	Test Data	**921 files ~ 1.45 hours**

METHODOLOGY

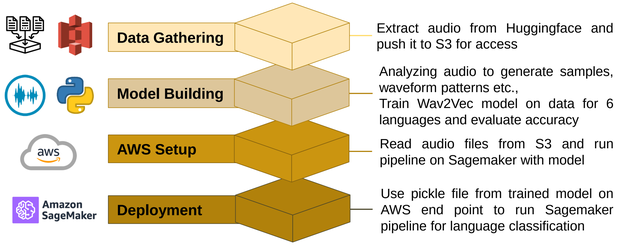


Fig 3. Project Outline

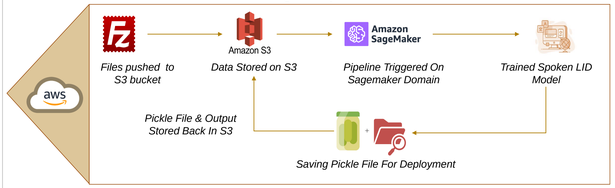


Fig 4. AWS Architecture for Deployment

MODEL BUILDING

True Label \ Predicted label	cak	es	kjb	quc	mam	kek
cak	113	0	1	0	0	1
es	0	189	1	0	0	0
kjb	0	0	89	0	0	2
quc	0	0	0	149	1	1
mam	0	0	0	0	166	1
kek	2	2	1	0	0	202

Fig 5. Wav2Vec Confusion Matrix

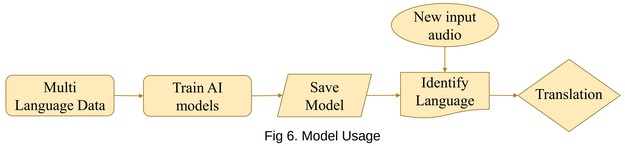


Fig 6. Model Usage

Areas For Improvement:
- ✓ Generalizing model to work with multiple speakers
- ✓ Making the model more robust to different accents
- ✓ Reducing the latency for language identification in the inference pipeline

DEPLOYMENT & LIFE CYCLE MANAGEMENT

Business Validation and Impact

Spoken Language Identification Model will help identify rare languages at various checkpoints and kiosks across globe, helping bridge language barriers and preserve disappearing

Fig 7. Estimated Impact due to Spoken Language Identification Model

"At least 6 million people in Guatemala, Mexico, Belize, and Honduras are Mayan speakers"

Approximately 200,000 people emigrate out of Guatemala each year

35% people in Guatemala speak Mayan Languages

Assuming 20-30% have negligible proficiency in English

40k – 60k people can be benefited from this model if we are able to deploy it at most of the checkpoints

Future Scope

Recognizing dialect is an important aspect when identifying spoken language and could further improve our model

Customizing AWS services as per deployment scale to avoid incurring unnecessary charges

Client Testimonial

"Training these new LID models will help better represent the world's languages and potentially could help preserve them."

ACKNOWLEDGEMENTS

We would like to thank our industry partner (SIL) for their guidance and support on this project as well as the Purdue MS BAIM program for partially funding this work.

Mitchell E. Daniels, Jr. School of Business