AUDIO-TO-TEXT ALIGNMENT FOR NON-DOMINANT LANGUAGES

SIL INTERNATIONAL

Mitchell E. Daniels, Jr. School of Business

The Team

Chinmay Gidwani

Email: chinmay.gidwani@gmail.com

LinkedIn: linkedin.com/in/chinmay-gidwani/

Keertana Madan

Email: madank@purdue.edu

LinkedIn: linkedin.com/in/keertanamadan/

Prathyusha Reddy

Email: pmidudhu@purdue.edu

LinkedIn: linkedin.com/in/prathyusha-reddy-616316219/

Kavyasri Jadala

Email: kjadala@purdue.edu

LinkedIn: linkedin.com/in/kavyasri-jadala/

Anupama M

Email: m1@purdue.edu

LinkedIn: linkedin.com/in/anupama-m

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

SIL's vision is to see people in communities flourishing using the languages they value most.

SCRIPTSOURCE
Writing Systems


World Language Map

SIL LEAD
Language, Education, and Development
Educational Empowerment

ISO 639-3
Language Codes

International Decade of Indigenous Languages 2022 - 2032
Language for Development

Ethnologue
Global Languages


Jonathan Hudlow
Senior Data Scientist


Joshua Nemecek
Data Scientist

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Story Flow

AGENDA

- 01 Business Problems & Benefits
- 02 Project Scope
- 03 Data Overview
- 04 Methodology

Purdue University | Mitchell E. Daniels, Jr. School of Business

Story Flow

AGENDA

05 Model Approach

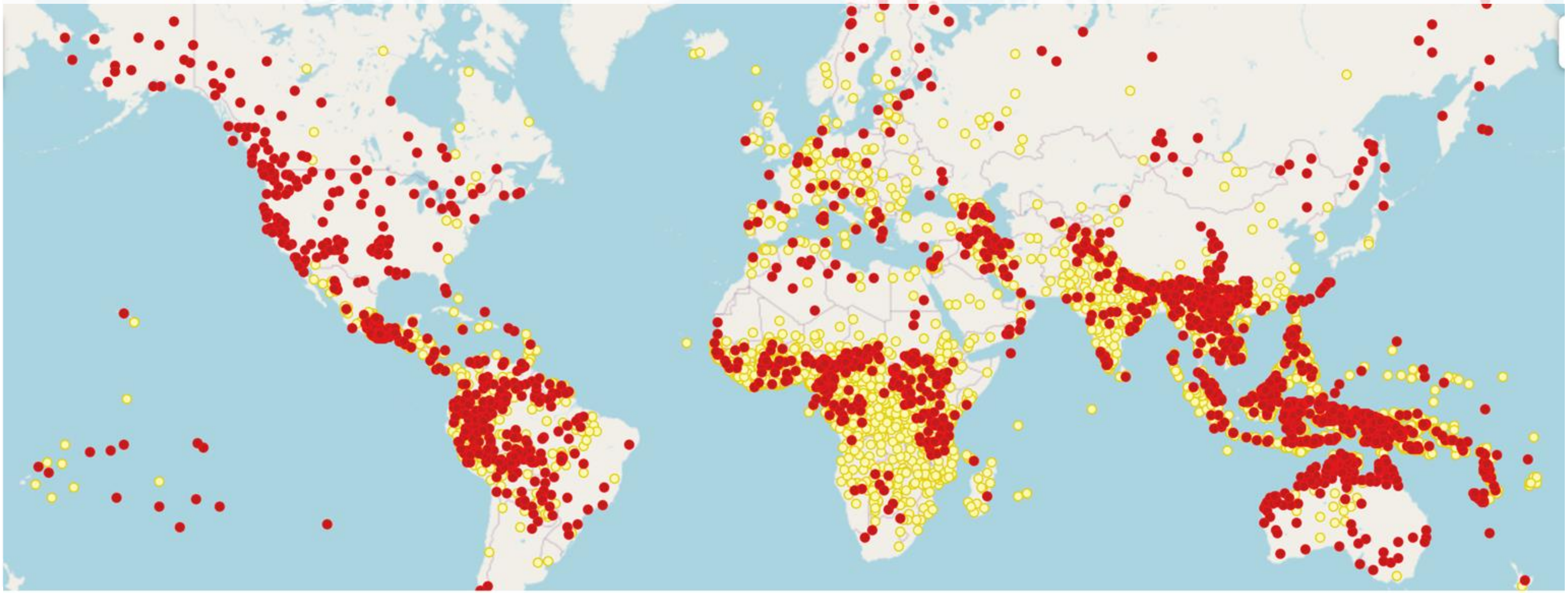
06 Results & Recommendations

07 Challenges

08 Future Scope

PURDUE UNIVERSITY
Mitchell E. Daniels, Jr. School of Business

Business Problem



Benefits

Direct Impacts

Translation Tools

Educational Material Creation

Customized Learning Tools

Broader Impacts

Enhanced Language Accessibility and Inclusivity

Improved Accuracy in Multilingual Contexts

Positive Social Impact

SIL

Project Scope

- Aim to establishing how much data is needed to train the alignment model to achieve high accuracy.
- Focusing on monotonic, monolingual audio from datasets like Mozilla Common Voice.
- Innovation lies in the method of generating phonemes from audio and aligning these with tokenized text, potentially requiring less data than conventional speech recognition technologies.

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Data Overview

Mozilla Common Voice
Recordings of various languages (monolingual)

Europarl-st
Multilingual European Parliament proceedings

Characteristics

- Language Diversity: Covers a 120+ wide range of languages, including low-resource languages.
- Format: Includes both audio and text transcriptions.

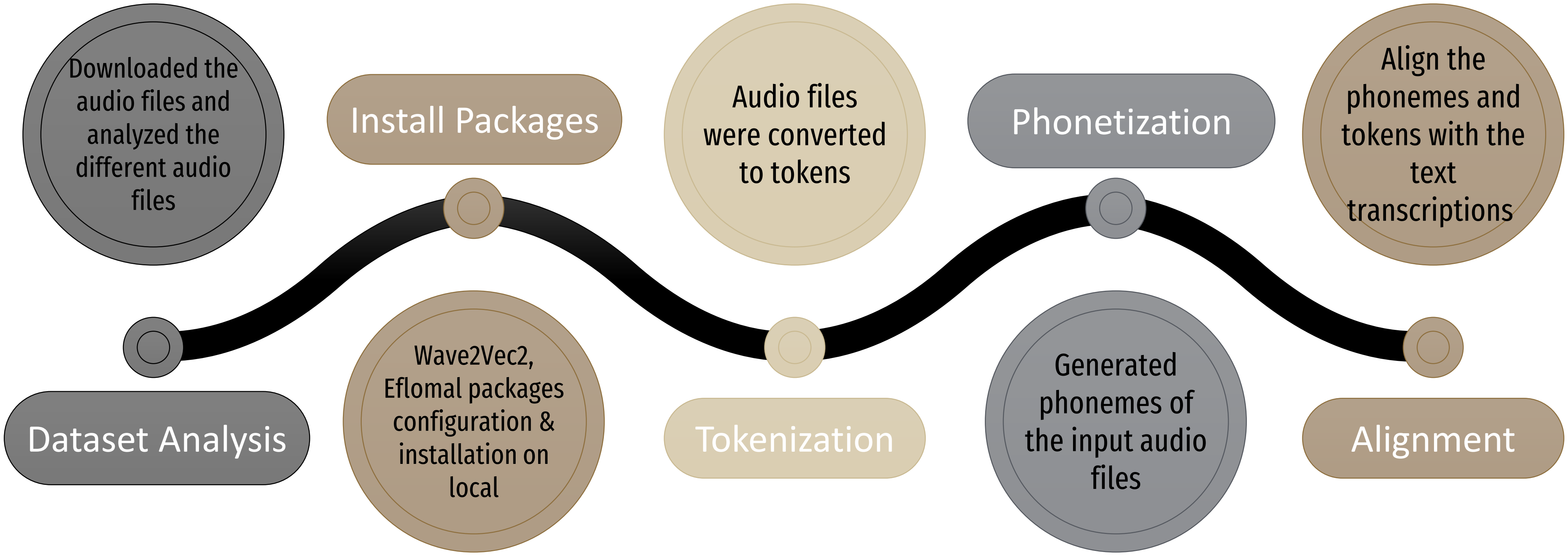
Purpose

- Audio-Text Alignment: To develop and test methodologies for aligning spoken words with written text.
- Cross-lingual Analysis: Facilitates experimentation in mono-lingual and cross-lingual contexts.

Languages

- English
- Hindi
- French

Methodology



SentencePiece

How SentencePiece transforms text data

Key Features

Lossless Tokenization

Sub-word Segmentation

Language Agnostic

Workflow

Training

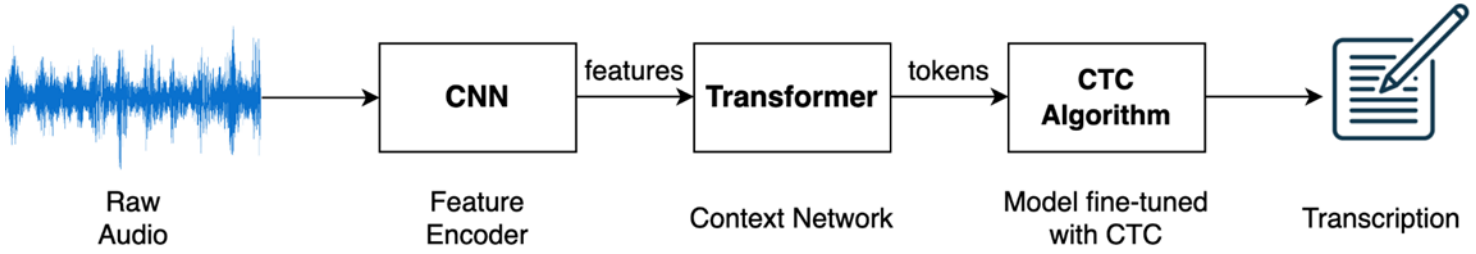
Tokenization

Detokenization

Advantage Over Traditional Tokenizers: Handles rare/unfamiliar words by breaking them into known segments, unlike traditional methods that might misinterpret or omit them.

Wav2Vec2

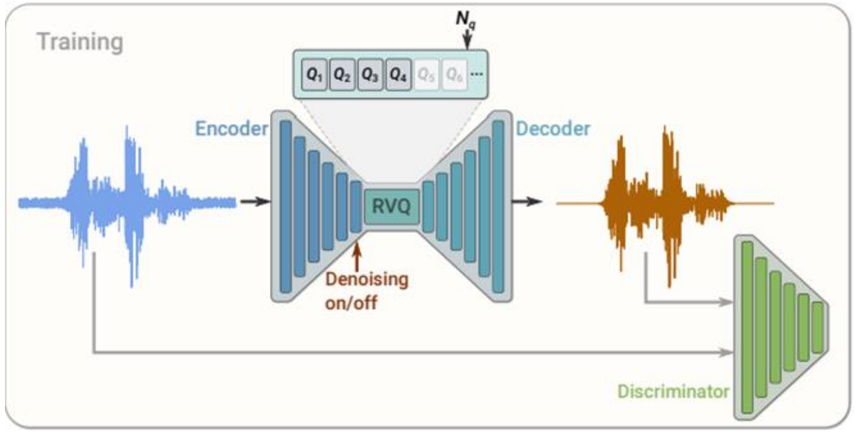
How Wav2Vec2Phoneme transforms Audio data into Phonemes



Advantage Over Traditional Tokenizers: Is language agnostic and converts audio data into sounds(phonetic) tokens.

SoundStream

How Soundstream transforms Audio data

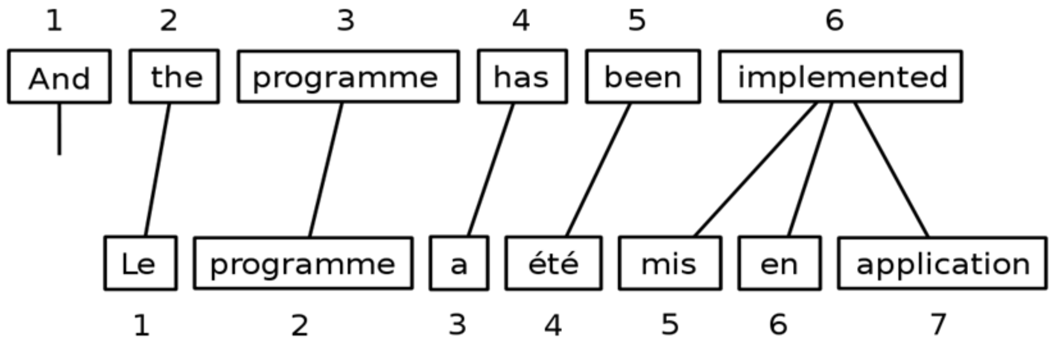


Advantage Over Wav2Vec2phoneme : Wave2vec2 phoneme typically produces about 300 output tokens, which contrasts with SoundStream's larger and adjustable token dictionary. However, the application of wave2vec2 phoneme is limited by its training scope, which does not encompass all languages. Consequently, we may need to consider SoundStream for broader linguistic coverage or demonstrate the effectiveness of wave2vec2 phoneme in languages it was not explicitly trained on.

Mitchell E. Daniels, Jr.
School of Business

Eflomal

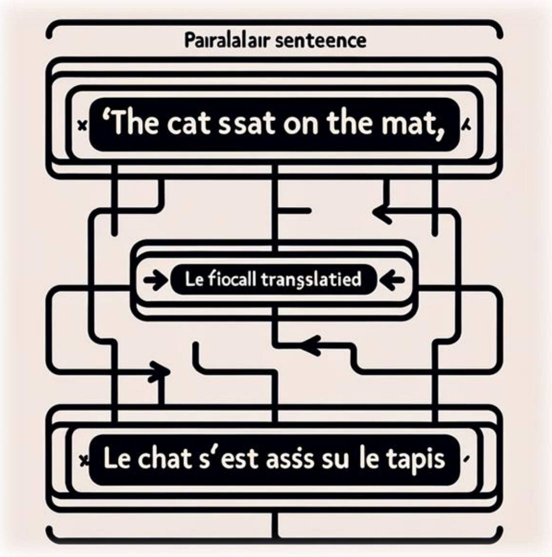
How Eflomal aligns text data

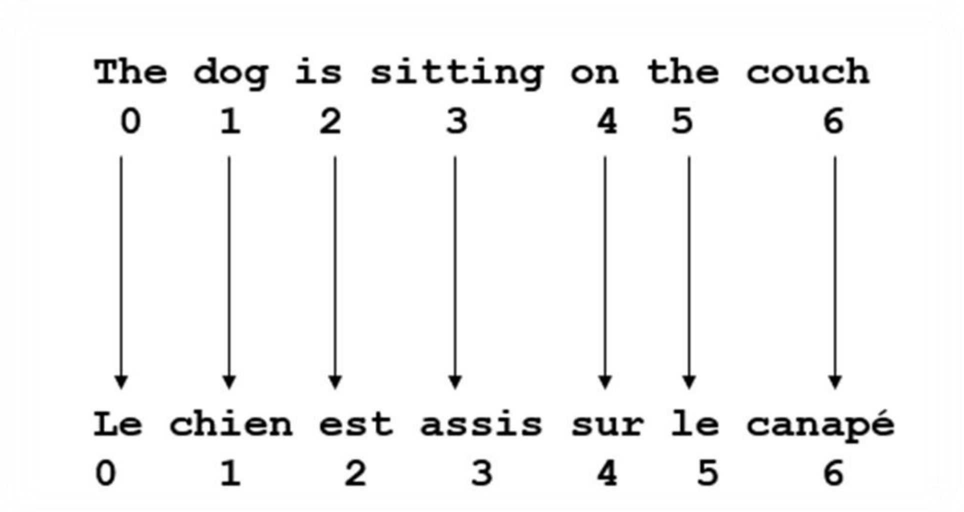


Advantage Over Traditional Tokenizers: Has the least alignment error rate as compared to other aligners available.

SimAlign

How Simalign aligns text data





Limitation compared to Eflomal: It has poorer performance compared to Eflomal for lesser known languages.

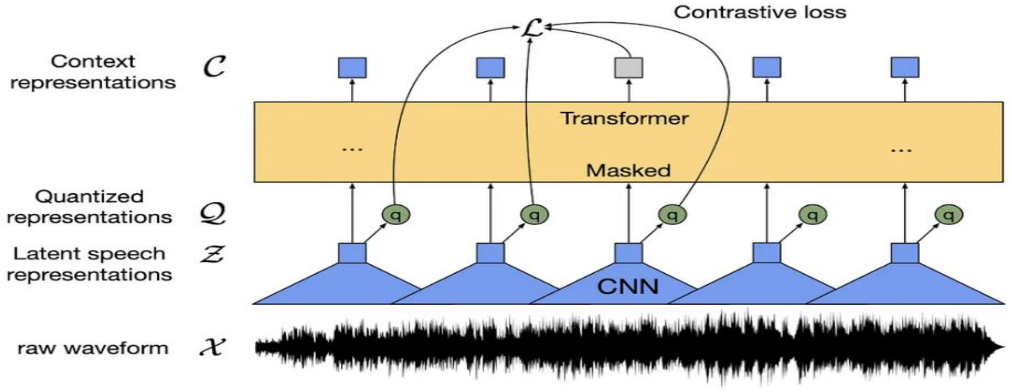
Model Deployment –Audio to transcript

STEP-1: Tokenization

Model	Trained model	Sample Lanugage	Duration
Wav2Vec 2.0 Base	XLSR-53	English	70 hrs
Wav2Vec 2.0 Large	XLSR-53	English	70 hrs
Wav2Vec 2.0 Large	LV-60	Hindi	21 hrs
Wav2Vec 2.0 Large	LV-60	French	16 hrs

Usage of pre-trained models for audio-to-text transcription

Spoken Language Waveform (16 Hz)



Components of Wav2Vec 2.0

A black cat .

Mitchell E. Daniels, Jr. School of Business

Transcriptions

Step.2: Wav2Vec2 Transcriptions

	path	sentence	file_name	transcription
0	common_voice_en_39586386.mp3	Well, that's not an indictment of Michael Jack...	common_voice_en_39586386.mp3	WELL THAT'S NOT AN INDICMENT OF MICHAEL JACKSO...
1	common_voice_en_39586336.mp3	The message is then read off in rows.	common_voice_en_39586336.mp3	THE MESSAGE IS THEN READ OFF IN ROSE
2	common_voice_en_39586337.mp3	In the days when the judges judged, there was ...	common_voice_en_39586337.mp3	IN THE DAYS WHEN THE JUDGES JUDGED THERE WAS F...
3	common_voice_en_39586338.mp3	Edwards was born in Shaw, Mississippi.	common_voice_en_39586338.mp3	EDWARDS WAS BORN IN SHAW MISSISSIPPI
4	common_voice_en_39586339.mp3	The qualifications for the award are determine...	common_voice_en_39586339.mp3	THE QUALIFICATIONS FOR THE AWARD ARE DETERMINE...

Language	File Size	Word Error Rate
German	1262.06 MB	18.5
Spanish	1262.08 MB	17.6
Polish	1262.08 MB	24.6
French	1262.09 MB	22.1
Portuguese	1262.1 MB	27.1
Hindi	1262.11 MB	25.2
Dutch	1262.12 MB	21.1

Model Deployment – Transcript Alignment

STEP-3: Alignment with transcripts

English transcript

test1.en

```
A black cat .
A yellow bird .
A white elephant .
```

Swedish transcript

test1.sv

```
En svart katt .
En gul fågel .
En vit elefant .
```

```
En svart katt . ||| A black cat .
En gul fågel . ||| A yellow bird .
En vit elefant . ||| A white elephant .
```

Bilingual Input Requirement:

- Requires input files in both languages for analysis.
- Eflomal processes these files to determine cross-language token mapping.

Forward Alignment

sv-en.fwd

```
0-0 1-1 2-2 3-3
0-0 2-2 3-3
0-0 1-1 2-2 3-3
```

Reverse Alignment

sv-en.rev

```
0-3 1-2 2-1 3-0
0-3 1-2 2-1 3-0
0-3 1-2 2-1 3-0
```

Forward Alignment:

- Sequential token mapping from source to target language.
- Begins at the start of the text, ensuring linear alignment progression.

Reverse Alignment:

- Token mapping from target back to source language.
- Starts from the end, offering a backward analysis perspective.

Generating Priors

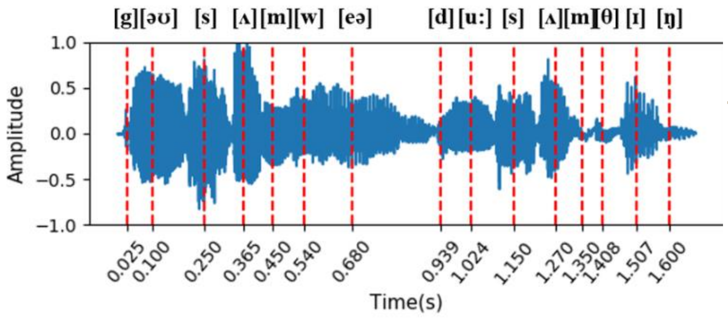
sv-en.priors

```
LEX	.	.	3
LEX	En	A	3
LEX	elefant	elephant	1
LEX	fågel	bird	1
LEX	katt	cat	1
LEX	svart	black	1
LEX	vit	white	1
FERF	.	1	3
FERF	En	1	3
```

- Utilizes statistical data to predict token alignments.
- Priors act as a reference guide, enhancing alignment accuracy by leveraging historical data.

Model Deployment

STEP-4: Phonemization



Example of Phoneme recognition Model: Wav2Vec2Phoneme

['aɪ m ɪ n t ɹ ɚ s t ɪ d oʊ n l i ɪ n ð ə p ɹ ɛ z ə n t t']

Transcript from Wav2Vec2Phoneme

Transcipt from Wav2Vec2Phoneme

	0
0	ð ə d ɛ f ə n ɪ t t t ɹ iː t m ə n t f ɚ h eɪ ...
1	p aɪ l z ʌ v s t oʊ n z m eɪ b iː s iː n æ t ð...
2	aɪ r ə m z k ə n b iː m ɪ k s t t ə g ɛ ð ɚ t ...
3	j ɪ h iː ə n ɐ d ɪ ɪ z ə n ɪ dʒ ɪ p ʃ ə n p l ...
4	n ɔ ɹ ɪ s g ɹ iː n p ɑː k ɪ z l oː k eɪ r ɪ d ...
5	b aɪ ð ɪ ɛ n d ʌ ð ə d ɛ k eɪ d ð ə h ɪ n ʒ m ...
6	ð ə m uː v w ʌ z ɐ s ɪ s t ɪ d æ n d f ə s ɪ l...
7	ð ə s k uː l ɪ v ɛ n tʃ uː l i b ɪ k eɪ m ð ə ...
8	ð ə f s ɪ l ɪ r i z h æ d θ ɹ iː k ɑː r ɪ dʒ ɪ...
9	ʌ n d ɚ s t æ n ə n k ɑː n t ɹ ɚ b j uː ʃ ə n ...

Ground truth transcripts

Mozilla Common Voice Transcipt

	0
0	The definitive treatment for Heyde's syndrome ...
1	Piles of stones may be seen at the partings of...
2	Items can also be mixed together to create dif...
3	Yahia Nader is an Egyptian player born in the ...
4	Norris Green Park is situated between Broad La...
5	By the turn of the decade, the hinged mitt bec...
6	This move was assisted and facilitated by the ...
7	The school eventually became the Technical Col...
8	The facility had three cottages for boys and o...
9	As Ford introduced new models, these were asse...

Forward Alignment

	0
0	0-0 4-1 14-2 18-3 23-4 29-5 33-6 36-7 44-8 51-...
1	3-0 5-1 8-2 12-3 13-4 16-5 18-6 20-7 24-8 28-9...
2	1-0 5-1 8-3 12-4 17-5 26-7 30-8 35-9
3	3-0 6-1 12-3 14-4 22-5 24-6 28-7 29-8 34-9 35-...
4	1-0 7-1 9-2 13-3 17-4 24-5 27-6 32-7 34-8 43-9...
5	1-0 2-1 8-4 13-5 15-6 20-7 20-8 24-9 33-10 38-...
6	0-0 4-1 5-2 8-3 15-4 25-5 30-6 31-7 34-8 44-9 ...
7	0-0 4-1 10-2 14-3 19-4 23-5 32-6 34-7 35-8 40-9
8	0-0 7-1 10-2 13-3 20-4 23-5 26-6 28-7 31-8 34-...
9	11-4 15-6 16-7 21-8

"cat" → "/k/, /æ/, /t/"

Audio file	Phonetic transcription (IPA)
./test_relecture_texte.wav	ʃapitʁ di də abɛse pəti kɔ̃t də ʒyl ləmɛtʁ ɑ̃ʁʒistʁe puʁ libʁivɔksɔʁg ibis dɑ̃ la bas kuʁ dœ̃ ʃato sə tʁuva paʁmi tut sɔʁt də volaj œ̃n ibis ʁɔz
./10179_11051_000021.flac	kɛl dɔmaʒ kə sə nə swa pa dy sykʁ supiʁa se foʁaz ɑ̃ pasɑ̃ sa lɑ̃g syʁ la vitʁ fɛ̃ dy ʃapitʁ kɛ̃z ɑ̃ʁʒistʁe paʁ sonjɛ̃ sɛt ɑ̃ʁʒistʁəmɑ̃ fɛ paʁti dy domɛn pyblik

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Challenges

Limited or no resources on packages

Packages we had to work with had little to no resources or implementation and support online

Environment Incompatibility

Installing the packages on the configured Sagemaker environment was incompatible with packages being used

Untapped Research Area

Navigating the complexity of accurate audio to text phonetization poses a significant challenge due vast number of languages and dialects

Phoneme Recognition Training

Identify phonemes across diverse languages presents a unique challenge, stemming from the absence of established training guidelines.

Wave2vec2Phoneme word boundaries

Phoneme model is not able to compute word boundaries in output

Mitchell E. Daniels, Jr. School of Business

Experiments and Validation

Experiments

#	Experiment	Dataset	Eflomal	SimAlign
1	Pre-trained transcription model (English)	3000 rows	29%	50%
2	Pre-trained transcription model + Transliteration (Hindi)	3000 rows	35%	38%
3	Language agnostic Phonetic model (English)	1000 rows	98%*	99%*
4	Language agnostic Phonetic model (English)	3000 rows	97%*	98%*

Validation Metric – Unalignment rate

unalignment_ratio = (total_unaligned / sentence_length) * 100

Where total_unaligned is the sum of number of words from target that are left unaligned to the source (missing words) and the number of words left unaligned from source to the target (the padding/extra words)

* For the Phonetic model, unalignment_ratio was calculated on the basis of number of unaligned words per sample after manual validation

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Future Scope

Future scope of this project lies in developing multilingual alignments .

Steps that can be taken to further this project:

Test the method on non-monotonic monolingual audio .

Explore cross-lingual alignment,

Gather data for target languages.

Simulate real-life scenarios

THANK YOU

Mitchell E. Daniels, Jr.
School of Business