Audio-to-Text Alignment

Anupama M, Chinmay Gidwani, Kavyasri Jadala, Keertana Madan, Prathyusha Reddy Midudhula, Matthew A. Lanham
Purdue University, Daniel's School of Business
m1@purdue.edu; cgidwani@purdue.edu; madaank@purdue.edu; kjadala@purdue.edu; pmidudhu@purdue.edu; lanhamm@purdue.edu

Business Problem Framing

Imagine unlocking the vast universe of digital content for every corner of the world, no matter how remote or how rarely their language is spoken.

25% of the world's people are left out because of language-related barriers.

In our partnership with SIL International, a non-profit organization, we are working to create a future where educational and informative content can speak directly to everyone in their native language, making learning and accessing information a seamless, inclusive experience.

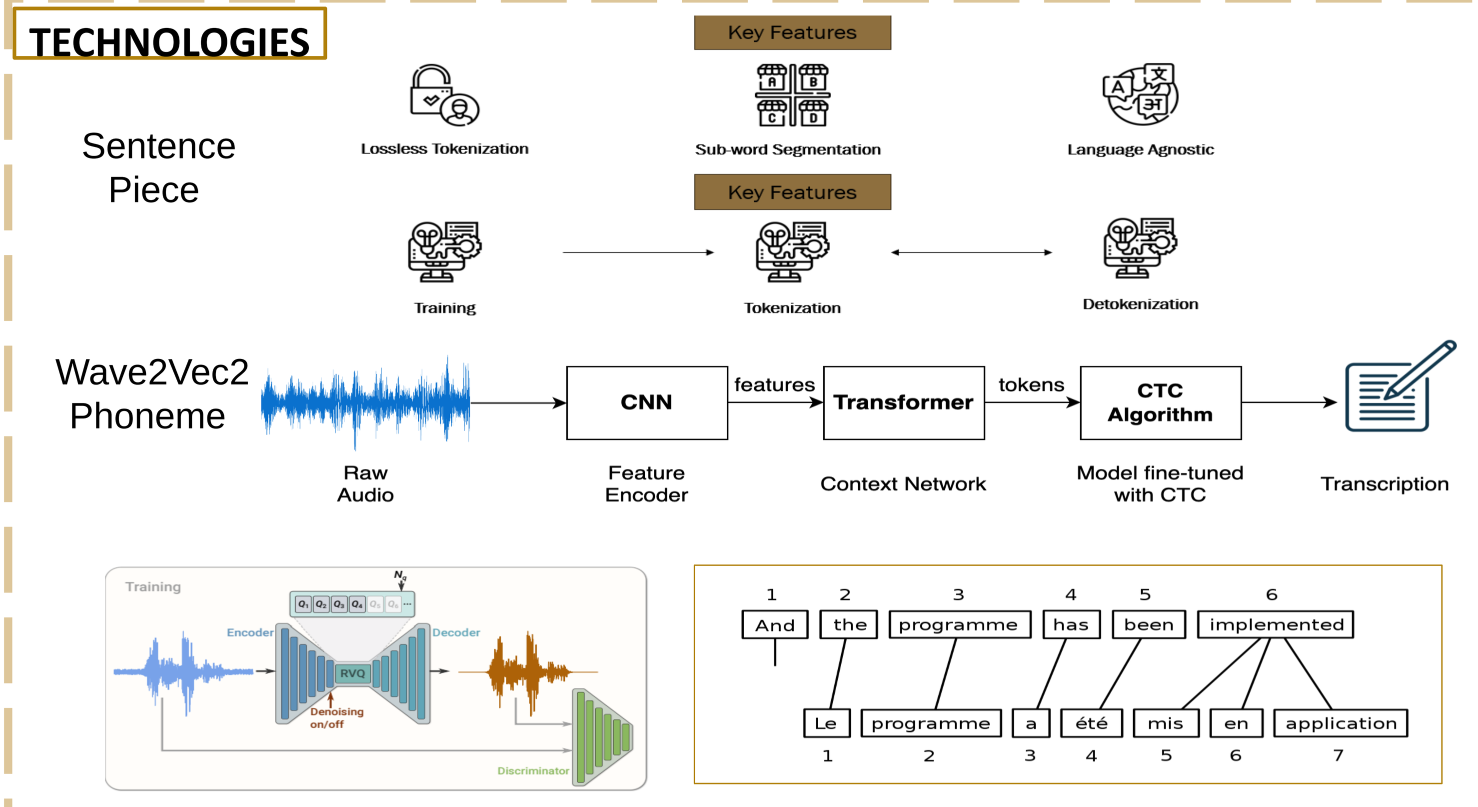
To answer this, our team is creating a language agnostic innovative system that automatically synchronizes audio and text across unseen languages to ensure they align without relying on speech recognition.

SIL's aim is to vastly simplify the creation of multilingual educational materials without the need for language experts using this solution.


Image generated from Dall-e

Now, the critical question we're exploring is: how much data is needed to teach this system effectively across languages that may share few similarities?

Analytical Problem Framing



We have used these methodologies to achieve significant audio to text alignment of three languages.

Data

Data from the open source, Mozilla Common Voice (audio + text translations) of 120+ languages .

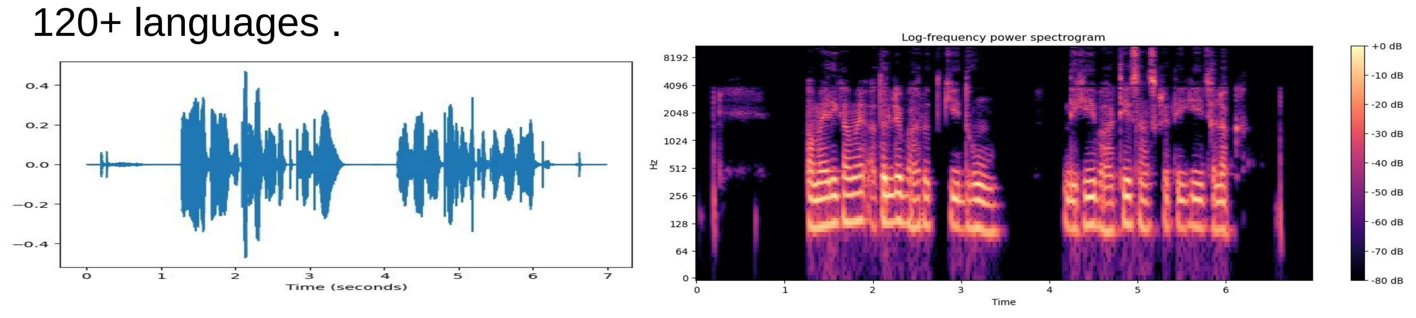


Fig 5. Waveform & Spectrogram analysis

Waveform : Loudness over time; Spectrogram : Constituent frequencies, Variation of pitch over time. Data Preprocessing: Extract Spectrograms, enhanced by normalization, compression, & noise reduction

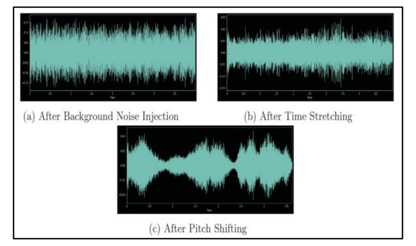
Datasets (Labeled- Audio)

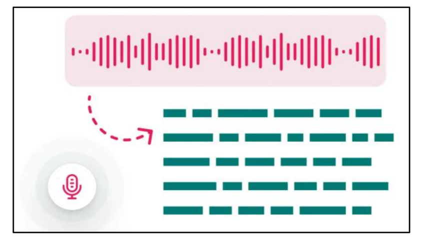
RAW DATA:

1. Mozilla Common Voice (Monolingual)
2. Europarl-st (Multilingual)

EXPERIMENTATION: INDUCING REAL WORLD INCONSISTENCIES

Missing Audio Simulation | **Missing Text Simulation**





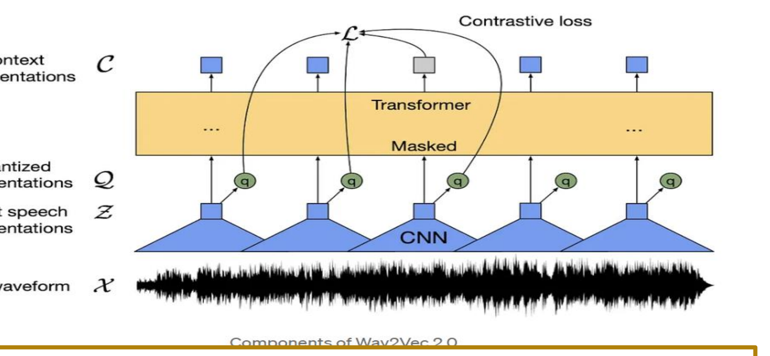
Introduce silences in random parts of audio files, matched with gaps in transcript to test performance w/real world data.

Methodology

TOKENIZATION

- The model advances with audio tokenization, employing Wave2vec2/Soundstream to produce phonemes using the processed audio in the previous step, which are then transformed into textual transcriptions. SentencePiece processes these transcriptions to generate text tokens.

Step-1 Audio Tokenization: Wav2Vec2Phoneme



Step-2 Text Tokenization: SentencePiece

Step-3 Fine tuned trained models for audio-to-text transcription (Sample)

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

Step-4 : Transcription Evaluation metric: Word error rate

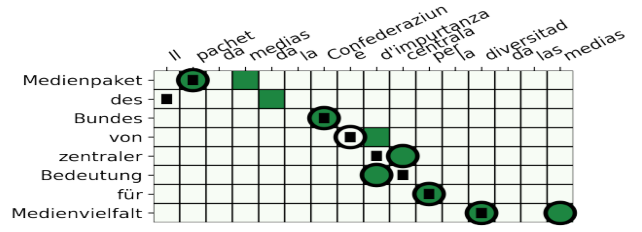
Language	File Size	Word Error Rate
German	1262.06 MB	18.5
Spanish	1262.08 MB	17.6
Polish	1262.08 MB	24.6
Italian	1262.09 MB	22.1
Portuguese	1262.1 MB	27.1
French	1262.11 MB	25.2
Dutch	1262.12 MB	21.1

Model Building

ALIGNMENT MODELS

Mono - Lingual Context

Eflomal
(Efficient Low memory aligner)



Cross - Lingual Context

Transformer Models

The outputs from tokenization are aligned using Eflomal to achieve a synchronized result. Training on the Common Voice dataset further refines this alignment and fortifies the understanding of the model.

Iteration and refinement

Validate Model accuracy
Assess and optimize model precision and efficacy

Re-train and optimize

Predictions

English transcript | **Swedish transcript**

test1.en
```
1 A black cat .
2 A yellow bird .
3 A white elephant .
```

test1.sv
```
1 En svart katt .
2 En gul fågel .
3 En vit elefant .
```

```
1 En svart katt . ||| A black cat .
2 En gul fågel . ||| A yellow bird .
3 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.

Model Deployment

Using eflomal we generate forward, reverse, and prior files which provides detailed mappings and alignment probabilities between languages.

Forward Alignment | **Reverse Alignment**

sv-en.fwd
```
1 0-0 1-1 2-2 3-3
2 0-0 2-2 3-3
3 0-0 1-1 2-2 3-3
```

sv-en.rev
```
1 0-3 1-2 2-1 3-0
2 0-3 1-2 2-1 3-0
3 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.

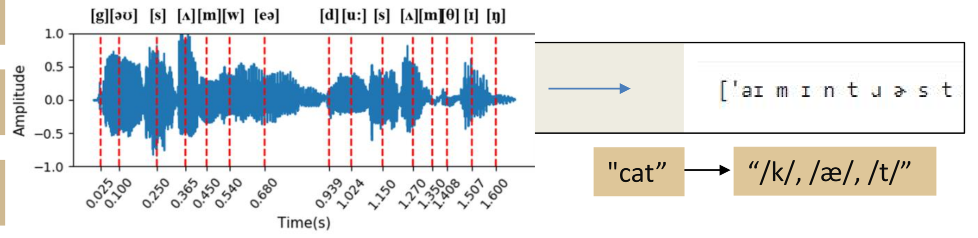
Reverse Alignment:
- Token mapping from target back to source language.
- Starts from the end, offering a backward analysis perspective.

Generating Priors

sv-en.priors
```
1 LEX . . 3
2 LEX En A 3
3 LEX elefant elephant 1
4 LEX fågel bird 1
5 LEX katt cat 1
6 LEX svart black 1
7 LEX vit white 1
8 FERF . 1 3
9 FERF En 1 3
```

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

Phonemization



Experiments with of Phoneme recognition Model: Wav2Vec2Phoneme

Alignment Evaluation metric: Alignment error rate

Alignment Error Rate (AER)						
Alignment models	De-En	Fr-En	Ro-En	Ja-En	Zh-En	En-Hi
fast_align	27	10.5	32.1	51.1	38.1	67.2
eflomal	22.6	8.2	25.1	47.5	28.7	46.7
Mgiza	20.6	5.9	26.4	48	35.1	55.7
Awesome-align	15.1	4.1	20.7	37.4	13.4	41.2
SimAlign	18.8	7.6	27.2	46.6	21.6	40.2
GIZA++	20.6	5.9	26.4	48	35.1	51.8
Efmaral	13.3	8.5	28.7	49	29.6	48.3

Comparing various alignment models, Eflomal worked best w.r.t Alignment error rate

Project Lifecycle Management

Initiation: Monolingual, Monotonic data identified. 3 languages picked as sample

Planning: Built an optimum and scalable architecture for deployment on multiple languages

Executing: **Current stage**, fine tuning our model and increasing the number of iterations to improve model quality

Monitoring: Continuously monitor model performance for anomalies during implementation. Training documentation

Future Scope: Expand to include non monotonic, cross-lingual and real-world application - educational videos and explore multimedia industry