Sanskrit Audio-To-Text Transcription

Team Members:

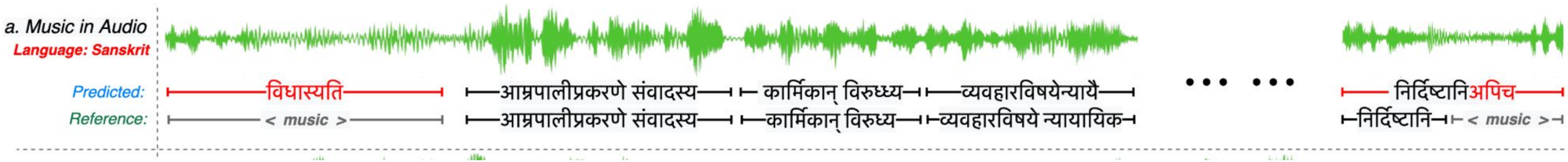
Venu Anupati,
Mounika Vempalli
Sai Saran Parasa

Problem Statement

- Taking one step forward to bridge the gap between ancient Sanskrit and modern English
- Addressing the challenge of limited technological support for Sanskrit
- Make Sanskrit's rich heritage accessible in today's world
- Developing a deep learning system for accurate Sanskrit Audio to text transcription

Dataset Collection

- Audio data from the News Services Division of All India Radio.
- Total duration of 27 hours with over 9700 samples with each audio ranging from 3 seconds to 35 seconds.
- Utilizing a subset of data from a study on mining audio and text pairs from public sources to improve Automatic Speech Recognition (ASR) systems for low-resource languages.



Data Understanding

Audio Data:

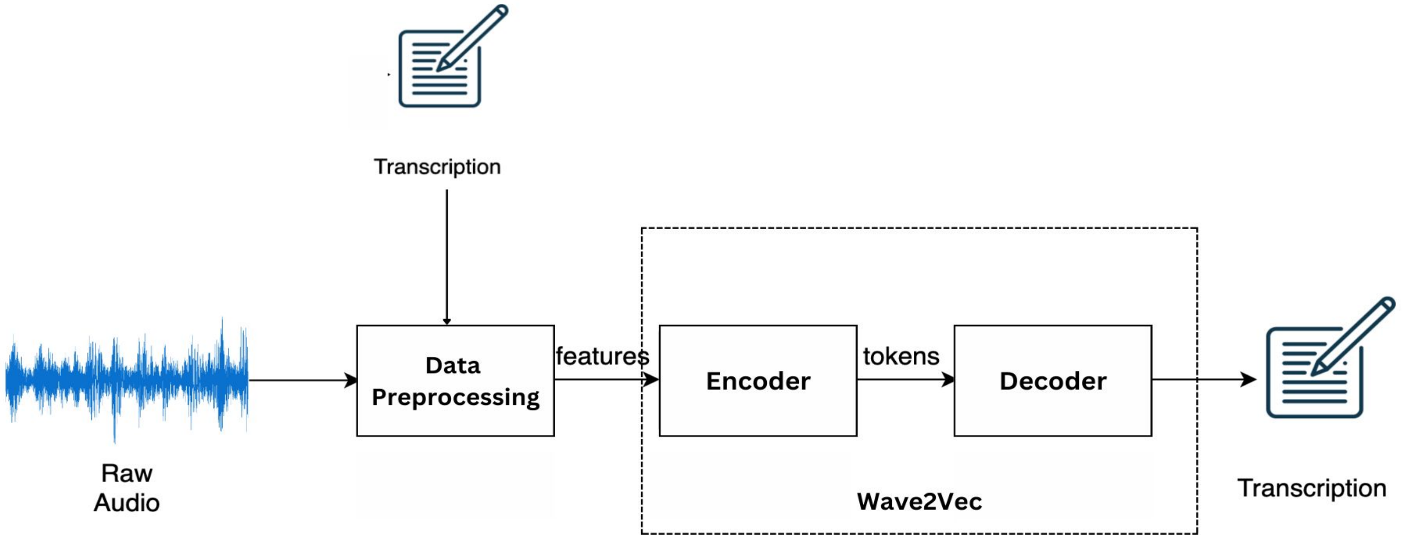
- Audio files are stored in separate folders, each corresponding to a news bulletin.
- The audio files are in wav format with a sampling rate of 16KHz.
- Filenames are structured with sentence IDs, such as sent_1.wav, for easy reference to their transcripts.

Data Understanding

Transcripts:

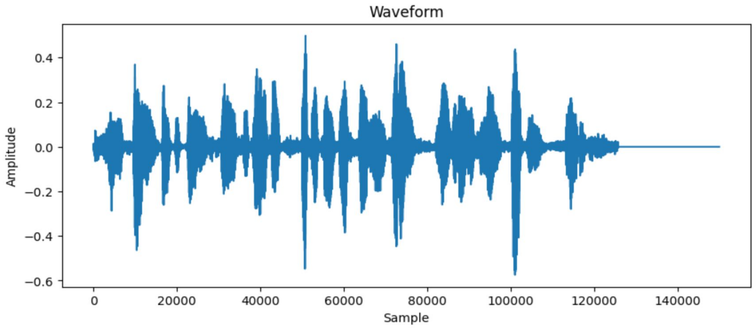
- Audio paths and Transcripts are stored in train.tsv, train.wrd
- The train.tsv file contains the relative path to an audio file and the number of frames in the audio, with an absolute path header.
- Train.wrd contains the word-level transcriptions corresponding to the audio files in train.tsv.

Architecture



Data Preprocessing

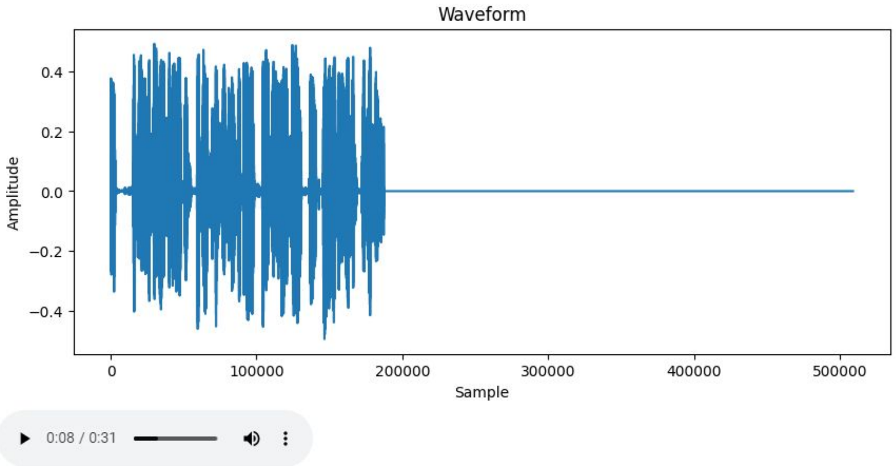
- **Audio Loading:** Loaded audio files and extract speech arrays and sample rates to prepare for analysis.
- **Transcript Tokenization:** Tokenized transcripts to convert them into input IDs for model processing..




```
Transcript: [' जीविंशति इति राष्ट्राणि वैश्विकार्थिकसमाह्वानाय ऐक्यमत्येन साहाय्याचरणाय प्रतिबद्धानि', ' प्रधानमन्त्रिणा स्ट्रेलियाया प्रधानमन्त्रिणा स्कटमरिसनवर्येण साकमपि वार्ताकृता यत्र क्रीडाखनन प्रौद्योगिकीरक्षासमुद्र
labels: ['| ज ी व ि ं श त ि | इ त ि | र ा ष ् ट ् र ा ण ि | व ै श ् व ि क ा र ् थ ि क स म ा ह ् व ा न ा य | ऐ क ् य म त ् य े न | स ा ह ा य ् य ा च र ण ा य |
Transcript: [' मूढै पाषाणखण्डेषु रत्नसंज्ञाविधीयते', ' राष्ट्रेण अद्य वरिष्ठभाजपानेत्रे प्राक्तन् विदेशमन्त्रिणे अरुणजेटलिने श्रद्धाञ्जलि समर्पितः', ' तस्मै अन्तिम श्रद्धाञ्जलिं प्रदातुं दलस्य नेतारः कार्यकर्ताश्च उपस्थिता अवर्तन्त
labels: ['| म ू ढ े | प ा ष ा ण ख ण ् ड े ष ु | र त ् न स ं ज ् ञ ा व ि ध ी य त े |', '| र ा ष ् ट ् र े ण | अ द ् य | व र ि ष ् ठ भ ा ज प ा न े त ् र े | प ् र ा
Transcript: [' प्रधानमन्त्री प्रधानमन्त्री नरेन्द्र मोदी अद्य स्वतंत्रता सेनान्या उपप्रधानमन्त्रिचरस्य च बाबू जगजीवनरामस्य द्वादशोत्तर शत तमीं जयन्तीम् उपलक्ष्य तस्मै श्रद्धाञ्जलिं समापर्यत्', ' कोविडनवदशाख्यं महामारीं 
labels: ['| प ् र ध ा न म न ् त ् र ी | प ् र ध ा न म न ् त ् र ी | न र े न ् द ् र | म ो द ी | अ द ् य | स ् व त ं त ् र त ा | स े न ा न ् य ा | उ प प ् र ध ा न म
Transcript: [' महाराष्ट्रहरियाणा विधानसभा निर्वाचनप्रक्रियान्तर्गतं अष्टाशीत्यधिकद्विशतमासनानाञ्कृते महाराष्ट्रे अथ च नवत्यासनेषु हरियाणा विधानासभानिर्वाचनेभ्यो सोमवासरे मतदानं अनुष्ठितमासीत्', ' जम्मू जम्मूकश्मीरे
labels: ['| म ह ा र ा ष ् ट ् र ह र ि य ा ण ा | व ि ध ा न स भ ा | न ि र ् व ा च न प ् र क ् र ि य ा न ् त र ् ग त ं | अ ष ् ट ा श ी त ् य ध ि क द ् व ि श त म ा
Transcript: [' जयशंकर शंघाईसहयोगसंघटनस्य महासचिवेन व्लादिमीर नोरो वर्येणाद्य नवदिल्ल्यां विदेशमन्त्रिणा एसजयशंकरेण मेलनं विहितम्', ' जयशंकरेण सदस्यदेशानां मध्ये पर्यटनं विवर्धयितुं शंघाई सहयोग संघट
labels: ['| ज य श ं क र | श ं घ ा ई स ह य ो ग स ं घ ट न स ् य | म ह ा स च ि व े न | व ् ल ा द ि म ी र | न ो र ो | व र ् य े ण ा द ् य | न व द ि ल ् ल ् य ा ं |
```

Data Preprocessing

- **Audio Padding**: Padded audio samples to a specified maximum length to ensure uniformity in the dataset
- **Label Padding:** Padded labels to a specified maximum label length for consistency in model training.
- **Tensor Conversion:** Converted input values to tensors needed for compatibility with deep learning frameworks.
- **Dataset Preparation**: Input values and labels are correctly formatted and ready for model training.



Wav2Vec2 Model

Model Type: Wav2Vec2ForCTC, based on the wav2vec framework developed by Facebook

Advanced Architecture: Utilizes a robust convolutional neural network to process raw audio data, extracting rich, contextual features without the need for manual feature engineering.

CTC for Alignment: Employs Connectionist Temporal Classification (CTC) to align speech inputs with their corresponding textual outputs, enabling effective transcription without requiring frame-wise alignment.

Self-Supervised Learning: Leverages self-supervised pre-training on vast amounts of unlabelled audio data, followed by fine-tuning on smaller annotated datasets, enhancing its ability to generalize across diverse linguistic contexts.

Optimized Performance: Designed with attention mechanisms and dropout strategies to enhance focus on relevant audio features and prevent overfitting, making it highly efficient for real-world applications.

Model Training

Training Configuration

- Directory: Outputs and models saved to "/content/drive/MyDrive/sanskrit/ABC"
- Batch Size: 2 per device during training and evaluation (Kept it 2 based on the resource available)
- Learning Rate: Starts at 1e-5, with warmup over 200 steps

Optimization Techniques

- Gradient Accumulation: 8 steps to effectively handle larger batch sizes
- Gradient Checkpointing: Enabled for memory efficiency
- Precision Training: FP16 enabled to accelerate computations

Model Training

Checkpointing

- Save Interval: Model and checkpoints saved every 250 steps
- Load Best Model: Automatically reloads the best model at the end of training

Training Execution

- Max Steps: Limited to 20000 to constrain the training period
- Trainer Setup: Utilizes Hugging Face's Trainer class for streamlined workflow
    - train_dataset for training
    - eval_dataset for validation

Hyperparameter Tuning

Batch size (per_device_train_batch_size, per_device_eval_batch_size): 2 (Kept as 2 to fit in with the resources available)

Learning rate: 1e-5

Training iterations: warmup_steps=200, max_steps =20000

Hardware optimization settings: gradient_checkpointing= Enabled

Evaluation settings:

- Evaluation_strategy: Steps = Steps determines that the evaluation will be performed at regular intervals based on the number of steps
- save_steps = 250
- Eval_steps = 250
- Load_best_model_at_end
- Metric_for_best_model: greater_wer_is_better=False

Model Evaluation

Evaluation Strategy: Conduct evaluations at regular steps

Metrics: Word Error Rate (WER) as the key metric

Best Model Criteria: Select model with lowest WER

As of now the WER is very high, meaning the predictions are not as per expectations.

Challenges:-

- High computational powered resources
- Limited Labeled Data
- Unique Phonetics or Distinct Sounds
- Language Structure (Grammar, Syntax etc)
- Computational Resources
- Language-Specific Adaptations

Future Work

- Try with batch size of 16 or 32 by putting high end resources, which can help the model to generalize better and give good predictions.
- This project can be expanded to transcribe from Sanskrit Audio to English text.
- Expanding the model to transcribe other less common languages, using insights from Sanskrit-to-English transcription to create a more versatile system.
- Create a system for live events or speech-to-text applications that transcribes speech in real-time for non english languages.

Conclusion

Our project is not just about transcribing ancient words into modern text, it's about preserving a language's legacy and making it accessible to generations to come

Thank You