Speech Recognition

Group: 4

-Vinal Gadhiya
-Jurin Vachhani
-Aashi Goyani

Project Overview

The goal of this project is to create a speech recognition system that bridges the gap between machine comprehension and human speech by applying cutting-edge deep learning techniques. With its ability to accurately translate spoken language into written text, the system offers a reliable solution for speech transcription in a variety of fields. Modern models like Whisper, Wav2Vec 2.0, and DeepSpeech are used to process audio inputs efficiently, extract useful information, and transcribe speech—even in difficult acoustic situations like noisy settings and different speech patterns.

Data Preparation

Model Implementation

Whisper (Open Ai)

Development Phase:

- Model: openai/whisper-small
- Encoder frozen to speed up training
- Feature Extraction: WhisperFeatureExtractor, WhisperTokenizer

Training Setup:

- Batch size = 4 (with gradient accumulation)
- Mixed precision (fp16), checkpoints
- Steps: 10 epochs (2200 steps)

Results

- WER: 34%
- Val Loss: 0.44

Wav2Vec 2.0

Whisper (Original) vs Whisper (Our Project):

Evaluation & Insights

Conclusion and Future Work

Thank You

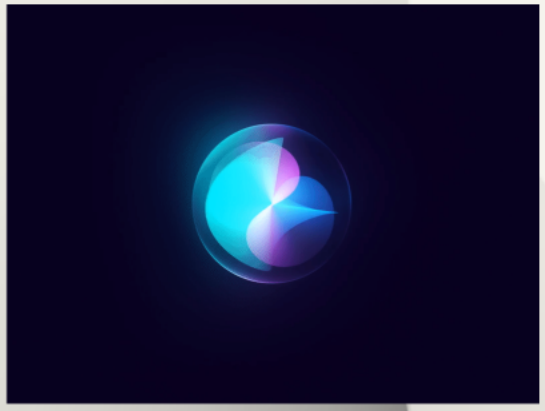
Speech Recognition

Speech-to-Text

Speech Recognition

Group: 4

-Vinal Gadhiya
-Jurin Vachhani
-Aashi Goyani



Project Overview

The goal of this project is to create a speech recognition system that bridges the gap between machine comprehension and human speech by applying cutting-edge deep learning techniques. With its ability to accurately translate spoken language into written text, the system offers a reliable solution for speech transcription in a variety of fields. Modern models like Whisper, Wav2Vec 2.0, and DeepSpeech are used to process audio inputs efficiently, extract useful information, and transcribe speech—even in difficult acoustic situations like noisy settings and different speech patterns.

Data Preparation

Acquired diverse audio samples to form a robust dataset for training and validation.

Processed the audio to enhance quality and ensure consistency for analysis.

Extracted key speech features to facilitate effective model training.

Acquired diverse audio samples to form a robust dataset for training and validation.

Processed the audio to enhance quality and ensure consistency for analysis.

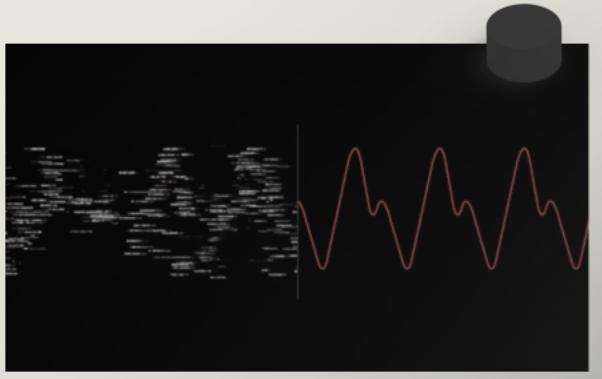
Extracted key speech features to facilitate effective model training.

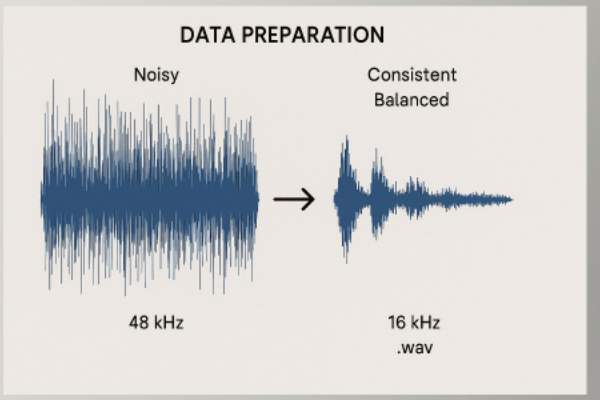
Data Preparation

Acquired diverse audio samples to form a robust dataset for training and validation.

Processed the audio to enhance quality and ensure consistency for analysis.

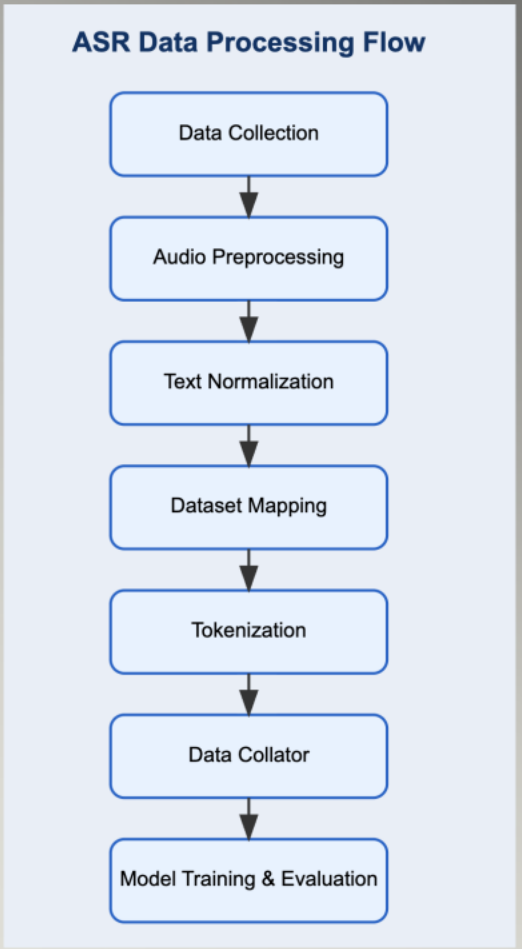
Extracted key speech features to facilitate effective model training.

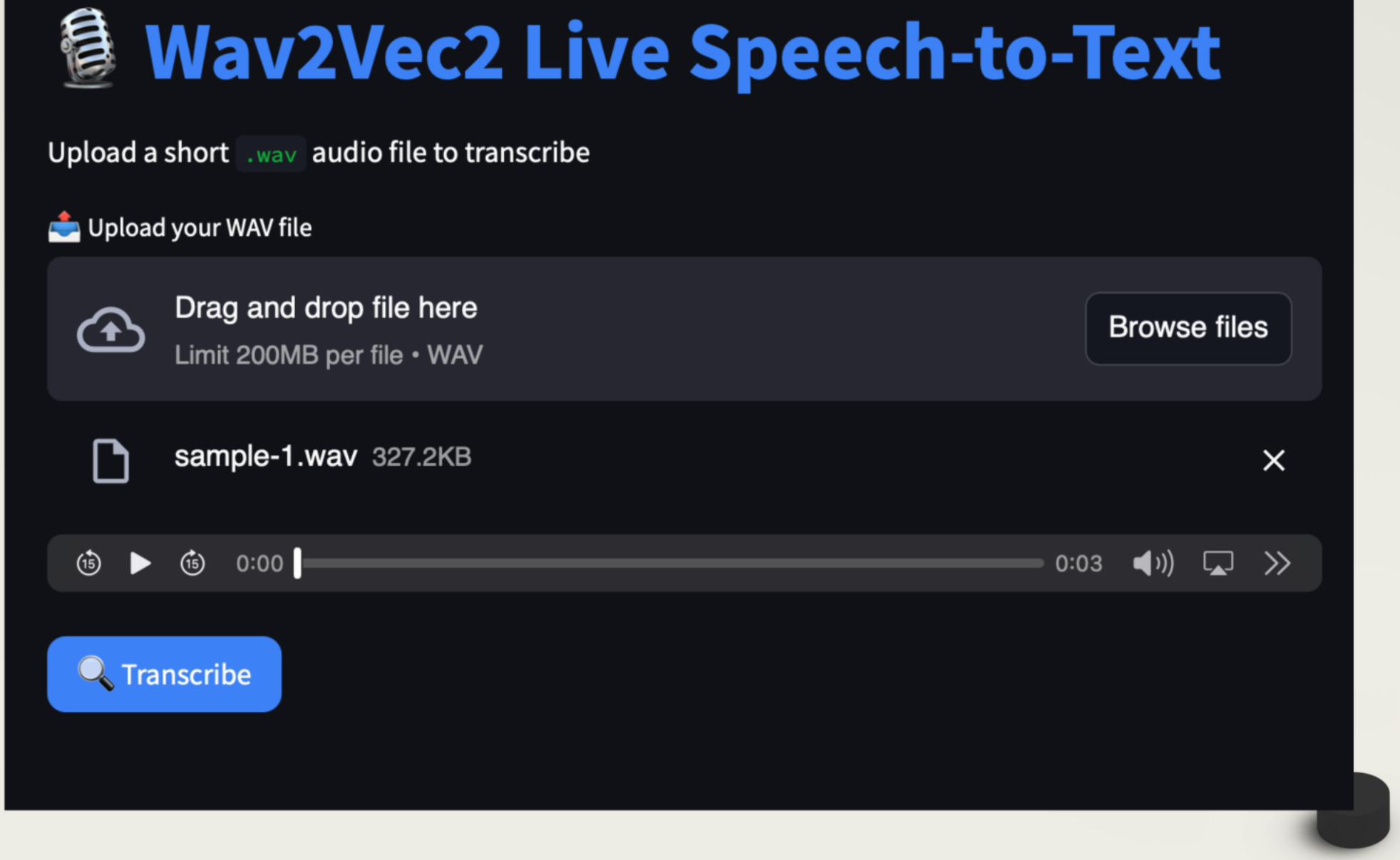




Model Implementation

- Researched and selected deep learning models:
  - Whisper (OpenAI)
  - Wav2Vec 2.0 (Meta)
  - DeepSpeech (excluded due to compatibility issues)
- Started implementing baseline model architectures
- Experimented with feature extraction techniques




Wav2Vec2 Live Speech-to-Text
Upload a short .wav audio file to transcribe
Upload your WAV file
Drag and drop file here
Limit 200MB per file • WAV
Browse files
sample-1.wav 327.2KB
0:00
0:03
Transcribe

Whisper (Open Ai)

Development Phase:

- Model: openai/whisper-small
- Encoder frozen to speed up training
- Feature Extraction: WhisperFeatureExtractor, WhisperTokenizer

Training Setup:

- Batch size = 4 (with gradient accumulation)
- Mixed precision (fp16), checkpoints
- Steps: 10 epochs (2200 steps)


```
"epoch": 9.777777777777779,
"grad_norm": 0.017008820548653603,
"learning_rate": 2.488888888888889e-07,
"loss": 0.0001,
"step": 2200


"epoch": 9.777777777777779,
"eval_loss": 0.09870120137929916,
"eval_runtime": 120.9928,
"eval_samples_per_second": 1.653,
"eval_steps_per_second": 0.207,
"eval_wer": 0.3448098663926002,
"step": 2200
```


Results

- WER: 34%
- Val Loss: 0.44

Wav2Vec 2.0

Development Phase

- Model: facebook/wav2vec2-base
- Data Preprocessing: Used preprocessed .csv and HuggingFace Dataset format
- Feature Extraction: AutoProcessor for audio + text

Training Setup:

- Batch size = 8 (grad accum = 4)
- Epochs = 10
- Learning Rate = 0.003
- CTC Loss used
- Mixed precision + gradient checkpointing

Results:

- WER: 10%
- Total CTC Loss: 125


```
"epoch": 9.777777777777779,
"grad_norm": 1173.87060546875,
"learning_rate": 1.5111111111111112e-06,
"loss": 50.7636,
"step": 2200


"epoch": 9.777777777777779,
"eval_loss": 125.71441650390625,
"eval_runtime": 48.8045,
"eval_samples_per_second": 4.098,
"eval_steps_per_second": 0.512,
"eval_wer": 0.10630722278738555,
"step": 2200
```

Whisper (Original) vs Whisper (Our Project):

Training Strategy:
• Original: Trained from scratch on 680k hours, multilingual
• Project: Fine-tuned on 2K English samples with encoder frozen

Resources:
• Original: 256 GPUs
• Project: Single GPU with FP16 + gradient checkpointing

Evaluation:
• Original: Active speech recognition across languages and noise
• Project: Word Error Rate (WER) on English dataset only

Robustness:
• Original: General-purpose multilingual model
• Project: Fine-tuned on cleaned, domain-specific Common Voice data

Aspect	Wav2Vec 2.0 (Original Paper)	Wav2Vec 2.0 (This Project)
Training Strategy	Pretrained with large batches and multi-GPU	Fine-tuned on 5k samples using single GPU
Resource Usage	FP32, fixed-length padding	FP16, dynamic batching, 30% memory reduction
Evaluation Focus	WER + PER (Phoneme Error Rate)	WER only (streamlined for focused comparison)
Robustness	Less focus on edge cases	Data validation to handle misaligned samples
Batching	Fixed-length batch processing	Custom collator + dynamic padding for variable-length inputs

Aspect	Wav2Vec 2.0 (Original Paper)	Wav2Vec 2.0 (This Project)
Training Strategy	Pretrained with large batches and multi-GPU	Fine-tuned on 5k samples using single GPU
Resource Usage	FP32, fixed-length padding	FP16, dynamic batching, 30% memory reduction
Evaluation Focus	WER + PER (Phoneme Error Rate)	WER only (streamlined for focused comparison)
Robustness	Less focus on edge cases	Data validation to handle misaligned samples
Batching	Fixed-length batch processing	Custom collator + dynamic padding for variable-length inputs

Evaluation & Insights

Evaluation Metrics:

- Word Error Rate (WER)
- CTC Loss (Wav2Vec)
- Validation Loss
- Cross-Entropy Loss(Whisper)

Observations:

- CTC Loss effectively optimized for Wav2Vec
- Model performance sensitive to preprocessing and batch size

Conclusion and Future Work

Whisper vs Wav2Vec 2.0:

Whisper:

- Moderate performance, Works without much fine-tuning (WER: 34%)
- Whisper is resource-efficient but lower accuracy

Wav2Vec 2.0:

- Higher accuracy, more compute-heavy (WER: 10%)
- Wav2Vec 2.0 achieves better performance at higher cost

Next Steps:

- Try larger datasets
- Deploy as a demo interface
- Explore multilingual datasets

Thank You

Speech Recognition

Group: 4

-Vinal Gadhiya
-Jurin Vachhani
-Aashi Goyani

Project Overview

The goal of this project is to create a speech recognition system that bridges the gap between machine comprehension and human speech by applying cutting-edge deep learning techniques. With its ability to accurately translate spoken language into written text, the system offers a reliable solution for speech transcription in a variety of fields. Modern models like Whisper, Wav2Vec 2.0, and DeepSpeech are used to process audio inputs efficiently, extract useful information, and transcribe speech—even in difficult acoustic situations like noisy settings and different speech patterns.

Data Preparation

Model Implementation

Whisper (Open Ai)

Development Phase:

- Model: openai/whisper-small
- Encoder frozen to speed up training
- Feature Extraction: WhisperFeatureExtractor, WhisperTokenizer

Training Setup:

- Batch size = 4 (with gradient accumulation)
- Mixed precision (fp16), checkpoints
- Steps: 10 epochs (2200 steps)

Results

- WER: 34%
- Val Loss: 0.44

Wav2Vec 2.0

Whisper (Original) vs Whisper (Our Project):

Evaluation & Insights

Conclusion and Future Work

Thank You

Speech Recognition

Speech-to-Text