Feeble Audio Based Transcript Generator

Addressing transcription challenges in noisy and feeble audio environments

Team Name: NLPeeps

*Team Members: **Aryash Srivastava** (22B1506), **Dhruv Garg** (22B1529)*

Guided by: Prof. Balamurugan

TA's: Mr. Rahul, Mr. Bheeshm

Problem Statement

Key Content:

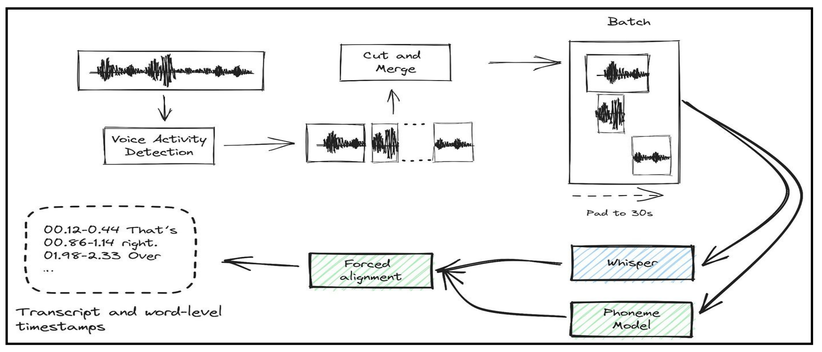
- ✔ Transcribing feeble and noisy audio with minimal syntax and information loss
- ✔ Identifying muted sections and accurately predicting missing content.

Techniques

- ✔ Speech-to-Text (STT).
- ✔ Audio Distortion: White noise, Gaussian noise.
- ✔ Spectrogram Analysis: Frequency-amplitude variation.

Objectives

- ✔ **Minimize transcription errors in low-quality audio.**
- ✔ **Handle muted segments using predictive modeling.**
- ✔ **Employ robust audio preprocessing for distortion handling**

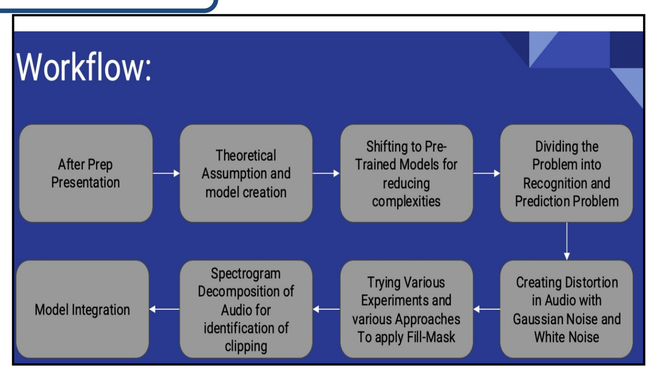


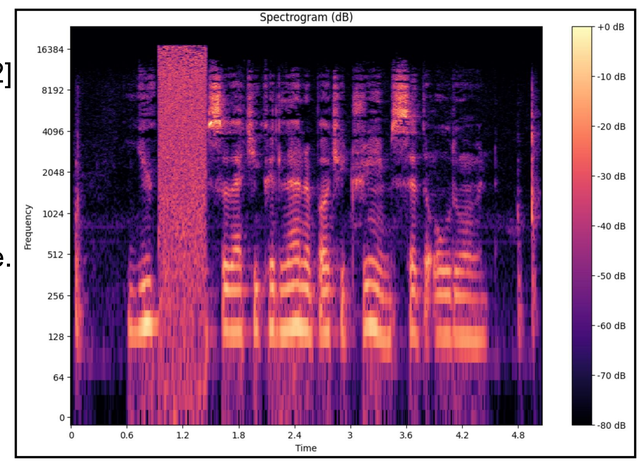
Models
Fill-Mask
huggingface.co/models

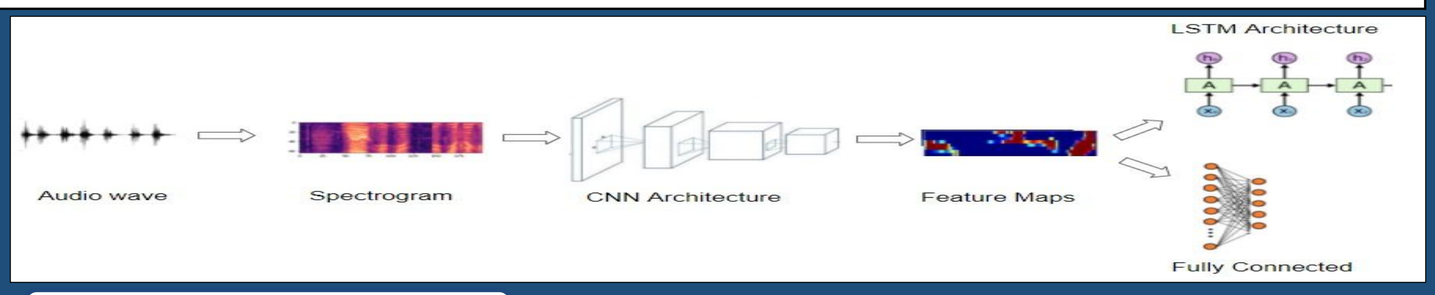
OpenAI Whisper

Workflow

1. **Theoretical Model:**
   - 1.1 Explored Hidden Markov Models, Lexicon Models.
   - 1.2 Speech prediction using n-grams.
2. **Transition to Pre-Trained Models:**
   - 2.2 Leveraged OpenAI Whisper for transcription.
3. **Audio Distortion Generation:**
   - 3.1 Added Gaussian noise ($N(0, \sigma^2)$), where, $\sigma \in [0.02, 0.2]$
   - 3.2 Replaced 0.2-0.5s segments with white noise.
4. **Muted Section Detection:**
   - 4.1 Spectrogram decomposition using amplitude variance.
5. **Final Integration:**
   - 5.1 Combined distortion handling, segmentation, and prediction into a unified model







Methodology

Data Preprocessing:

- ✔ Dataset: Mozilla "Common Voice" (200,000 samples, 3-7s duration).
- ✔ Clipping and noise augmentation using librosa and numpy.random.

Base Models:

- ✔ Distortion Generator: Adds noise to simulate real-world conditions.
- ✔ Spectral Differentiator: Segments audio based on silence and amplitude.
- ✔ Prediction Model: Uses masked language modeling to predict missing words.

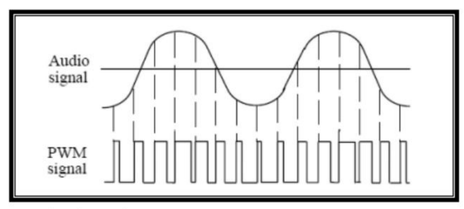
Integration:

- ✔ Combined models into a unified pipeline.
- ✔ Final predictor automates audio distortion handling and transcription.

Equations

Noise Addition:

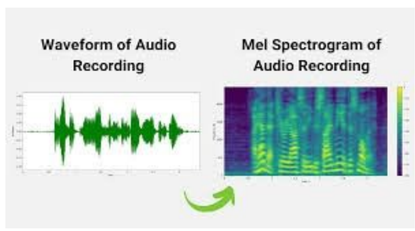
$y(t)=x(t)+\eta(t)$, where $\eta(t) \sim N(0,\sigma^2)$



Spectrogram:

$S(f,t)=|STFT(x(t))|^2$

where STFT is the Short-Time Fourier Transform



Results

Dataset Testing

Input: *Then I got a hold of some **dough** and went goofy.*

Predicted: *Then got a hold of some **joe** and went goofy.*

Metric	Baseline Model	Proposed Model
Accuracy (%)	65%	85%
Noise Resilience (%)	50%	78%

Real-Time Testing

Input: Live audio: *Sun from East*

Predicted: *Sun rises from East*

```
text1: Sun
text2:  from east of that building
Masked: Sun <mask> from east of that building
Sunrise from east of that building
```

Challenges & Limitations

Challenges:

- Handling extreme noise levels.
- Detecting and predicting for multiple muted sections.

Identified Inefficiencies:

- Only first pause is handled.
- No built-in denoising
- Language mismatch in predictions.

Whisper model architecture