Audio Extraction

Submitted to: Sir Adeel Nisar
Submitted by: Usama Laeeq | Muhammad Muaz Ashraf

Abstract: Key Findings

Develop a model to extract accurate text from audio using two techniques: fine-tuning and custom ASR.

Technique 1: Fine-Tuning

Fine-tune Whisper model and extract key info with Qwen2 model.

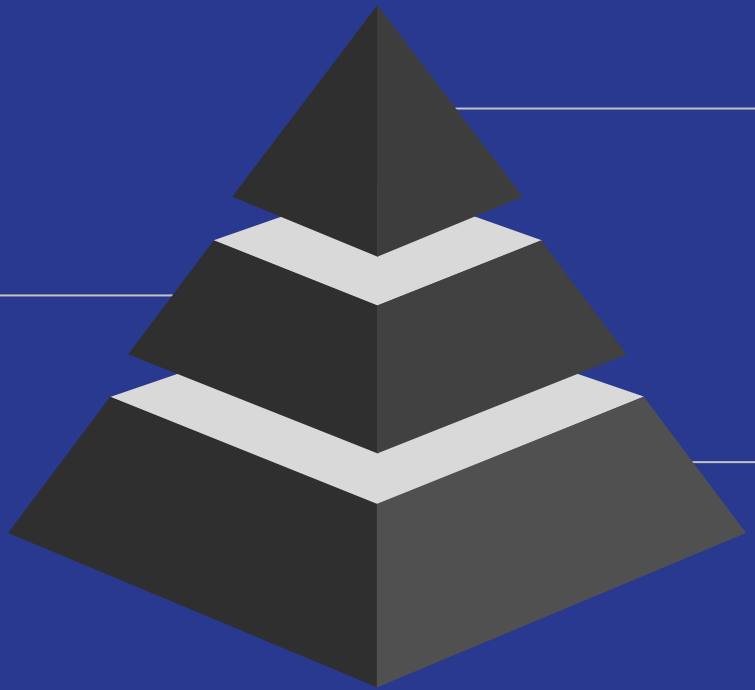
Technique 2: Custom ASR

Build ASR from scratch using CNN encoder and GRU decoder with Attention

Introduction to ASR

- **What is ASR:** Automatic Speech Recognition converts spoken language into text. Crucial in voice assistants like Siri, Alexa, and meeting transcription tools.
- **Key Challenges:** Accurate transcription despite noise, varying accents, speech speed, and pronunciation. Real-world audio is often messy.
- **Motivation:** Inspired by the use of ASR in daily tools and applications. Aim was to understand, compare, and build ASR systems.
- **Two Approaches:** Fine-tuning a pre-trained model vs. building a new model from scratch to solve the same task.

Technique 1(Fine-tuning)



1

Feature Extractor

Converts audio waves to mel spectrograms

2

Tokenizer

Converts spectrograms into numerical representations

3

Transformer Architecture

Whispers encoder decoder predicts text in parallel outperforming RNN

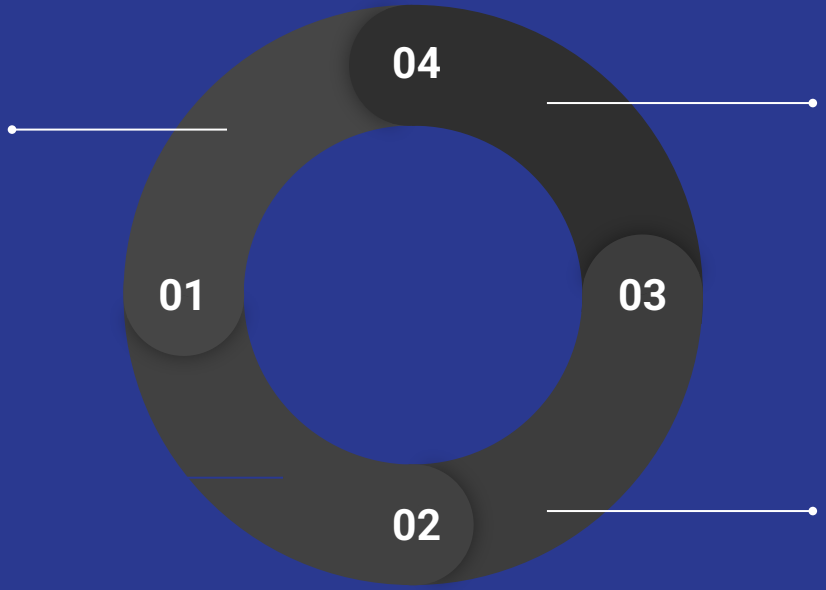
Technique 2 (Custom ASR)

CNN Encoder

Extracts features from mel spectrograms

GRU Decoder

Generates texts simpler and less expensive than LTSM



Attention Mechanism

Help decoder focus on specific audio segments

Teacher Forcing

Ensures correct decoder operation during training

Relevant Work

- **Whisper-Tiny (Fine-tuned):** Used on LibriSpeech (clean); achieved 6.6% WER. Fine-tuned using Seq2SeqTrainer.
- **Streaming ASR:** WeNet toolkit with Unified Two-Pass (U2) and causal attention. Effective real-time transcription.
- **Code Switching:** Adapted Whisper on SEAME and ASRU2019. Small data (1–10 hrs) yielded significant performance boost.
- **Aviation ASR:** Distil-Whisper model on air traffic control data (70 hrs); achieved 3.86% WER using LoRA fine-tuning.

Experiments and Results

- **Dataset Description:** LibriSpeech ASR (100 hrs subset, 28,539 samples). High-quality male/female voices. Sampling rate: 16kHz.
- **Data Settings:** 70% training, 20% validation, 10% testing. Seed = 42. Evaluated using WER metric.
- **Fine-Tuning Results:** Achieved 3.95% WER – significant improvement over 5% baseline.
- **Custom ASR Results:** 1.63% WER but considered unreliable due to unstable training conditions.

Discussion

- **Custom Model Issues:** Despite low WER (1.63%), custom model suffered from unstable training loops, limiting reliability.
- **Fine-Tuning Advantage:** Transformer-based Whisper model showed better learning, consistency, and robustness.
- **Training Constraints:** Limited to 10 epochs and smaller datasets due to hardware limitations. Optimizer state loss due to interruptions.
- **Strategic Takeaway:** Pretrained models significantly reduce effort and offer better out-of-the-box accuracy in real-world tasks.

Conclusion

- **Main Contributions:** Implemented fine-tuning of OpenAI Whisper and developed custom CNN-GRU-Attention ASR for comparison.
- **Conclusion:** Fine-tuned model more reliable with 3.95% WER. Custom ASR less stable despite 1.63% WER.
- **Future Plans:** Train on larger datasets with more epochs and stable GPU resources to improve training reliability.
- **Research Outlook:** Explore use of CTC Loss and advanced attention mechanisms to boost custom model performance.