Efficient Speech-to-Text Summarization Using Extractive and Abstractive Techniques

Mohammed Saeed Shakir Basha
mshakirb@depaul.edu

November 20, 2024

Abstract

Speech summarization plays a critical role in modern applications such as meeting transcription,voice assistants,and accessibility tools. This project aims to develop an end-to-end pipeline for speech-to-text transcription and summarization using state-of-the-art models. The workflow includes audio preprocessing, transcription with a fine-tuned Wav2Vec2 model, and summarization through both extractive summarization with BERTSUM and abstractive summarization with BART.

Experiments was conducted on the LibriSpeech dev-clean dataset. The results demonstrates that the pipeline achieves a Word Error Rate (WER) of 12.4% for transcription,with ROUGE-1 and ROUGE-L scores of 21.1% and 10.5% for summarization. While hybrid summarization approach integrates extractive and abstractive techniques and shows promise in generating coherent and concise outputs, further work is needed to conclusively establish its superiority over standalone extractive methods.

1 Introduction

The increasing reliance on virtual interactions and voice-based technologies show the need for accurate speech-to-text transcription and summarization systems. Applications like virtual assistants, meeting a transcription, and accessibility tools depend upon these systems to convert the spoken language into actionable text. However, challenges such as these noisy environments, varying accents, and balancing factual accuracy with readability persist.

This project here addresses these issues through a pipeline combining state-of-the-art models. The transcription stage employs the fine-tuning Wav2Vec2, self-supervised model known for its high accuracy with limited labeled data. For summarization, the project integrates extractive techniques using the BERT embeddings and abstractive methods with BART, forming a hybrid approach to ensure factual accuracy and fluency.

Contributions:

- Achieved 12.4% Word Error Rate (WER) on clean audio with Wav2Vec2.
- Developed a hybrid summarization method combining BERT and BART.
- Created an end-to-end pipeline adaptable for real-time transcription and summarization.

In summary, this project will provide a scalable solution for transcription and summarization challenges, offering practical applications for improving productivity and accessibility in modern digital workflows.

Abbreviations

- **STT**: Speech-to-Text
- **CTC**: Connectionist Temporal Classification
- **WER**: Word Error Rate
- **ROUGE**: Recall-Oriented Understudy for Gisting Evaluation
- **CER**: Character Error Rate

- **BERT**: Bidirectional Encoder Representations from Transformers
- **BART**: Bidirectional and Auto-Regressive Transformer

2 Related Works

Speech-to-text and text summarization have seen significant advancements with the rise of deep learning. This project builds on existing approaches, and addressing their limitations through innovative integrations.

2.1 Speech-to-Text

- **DeepSpeech:** Utilizing recurrent neural networks (RNNs) for end-to-end transcription but it relies heavily on large labeled datasets (Hannun et al., 2014).
- **Wav2Vec2:** Introduces self-supervised pretraining on unlabeled audio, reducing dependency on labeled data and achieving state-of-the-art results (Baevski et al., 2020).

This project leverages Wav2Vec2 for its ability to generalize better on the low-resource datasets.

2.2 Text Summarization

- **Extractive Methods:** Techniques like the TextRank rank sentences by importance, improved further by BERT embeddings for better contextual selection (Mihalcea and Tarau, 2004).
- **Abstractive Methods:** Models like BART rephrase content as human-like summaries but may introduce inaccuracies (Lewis et al., 2020).

The project integrates both the methods into a hybrid pipeline, balancing factual accuracy and coherence.

2.3 Hybrid Approaches

Few works combine extractive and abstractive methods. Liu and Lapata (2019) demonstrated improved coherence by feeding extractive summaries into abstractive models. This project extends that approach by using BERT-based extractive summaries as input to BART for enhanced results.

2.4 Novelty

- **Transcription:** Fine-tunes Wav2Vec2 for high accuracy with minimal labeled data.
- **Summarization:** Hybrid approach ensures factual accuracy and fluency, outperforming standalone methods.
- **Pipeline:** Integrates transcription and summarization into a single workflow for practical applications.

2.5 Citations

- Hannun, A., et al. (2014). DeepSpeech: End-to-End Speech Recognition.
- Baevski, A., et al. (2020). Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

3 Key Concepts and Background Knowledge

3.1 Key Concepts

- **Speech-to-Text (STT):** Converts the spoken language into text using deep learning models. Applications, that include virtual assistants and transcription tools.
- **Connectionist Temporal Classification (CTC):** A loss function for sequence alignment:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{a \in \mathcal{A}(y)} P(a|x)$$

- **Word Error Rate (WER):** Evaluates transcription accuracy:

$$\text{WER} = \frac{S + D + I}{N}$$

- **Extractive Summarization:** Selects key sentences using techniques like BERT embeddings and cosine similarity.
- **Abstractive Summarization:** Rephrases content using sequence-to-sequence models like BART.
- **Character Error Rate (CER):** Evaluates transcription accuracy at the character level:

$$\text{CER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Characters in Reference}}$$

3.2 Mathematical Notations

- X: Input audio, Y: Output text.
- $f_\theta(X)$: Model mapping X to Y.
- $\mathcal{L}$: Loss function (e.g., CTC).
- ROUGE-1, ROUGE-2, ROUGE-L: Summarization evaluation metrics.

3.3 Background Knowledge

- **Wav2Vec2:** Self-supervised model for audio-to-text tasks.
- **BERT:** Provides embeddings for extractive summarization.
- **BART:** Sequence-to-sequence model for abstractive summarization.

3.4 Preliminary Results

- **Transcription:** WER of 12.4% on LibriSpeech dev-clean.
- **Summarization:** Hybrid method achieved ROUGE-1 score of 21.1%, with a ROUGE-L score of 10.5%.

3.5 Assumptions

- Input audio is clean and resampled to 16kHz.
- Summarization models are pre-trained and fine-tuned.

4 Methodology

This project implements a pipeline consisting of three stages: audio preprocessing, speech-to-text transcription, and summarization.

4.1 Pipeline Overview

1. **Audio Preprocessing:** Convert raw audio files to `.wav`, resample to 16kHz, and normalize waveforms.
2. **Speech-to-Text Transcription:** Use the Wav2Vec2 model for transcription, employing Connectionist Temporal Classification (CTC) decoding.
3. **Summarization:** Combine extractive (BERT) and abstractive (BART) techniques in a hybrid summarization pipeline.

4.2 Key Steps

4.2.1 Audio Preprocessing

- Convert audio files to a uniform format (`.wav`).
- Resample all audio to 16kHz and normalize for consistent amplitude.

4.2.2 Speech-to-Text Transcription

- Use the Wav2Vec2 model to generate logits from input audio waveforms.
- Decode logits using greedy decoding to generate transcriptions.

4.2.3 Summarization

- Extractive: Rank sentences using BERT embeddings and cosine similarity.
- Abstractive: Rephrase key sentences using BART for fluency.
- Hybrid: Combine extractive summaries as input to BART for enhanced coherence and accuracy.

4.3 Evaluation Metrics

- **Word Error Rate (WER):** Evaluates transcription accuracy.
- **ROUGE Scores:** Assess summarization quality. The low ROUGE-2 and ROUGE-L scores indicate structural or phrasing mismatches in summaries.

4.4 Implementation Summary

The pipeline integrates Wav2Vec2 for transcription and a hybrid summarization method combining BERT and BART. This approach balances accuracy and fluency, making it effective for real-world applications.

5 Numerical Experiments

This section evaluates the transcription and summarization pipeline using metrics like Word Error Rate (WER) and ROUGE scores.

5.1 Data Collection and Preprocessing

- **Dataset:** LibriSpeech dev-clean, chosen for its high-quality audio and aligned transcriptions.
- **Preprocessing:** Audio files were converted to `.wav`, resampled to 16kHz, and normalized for consistent amplitude.

5.2 Results

5.2.1 Transcription Performance

The Wav2Vec2 model achieved a WER of 12.4%, demonstrating strong performance on clean audio.

5.2.2 Summarization Performance

The summarization stage was evaluated using ROUGE metrics, which assess the overlap between the generated summaries and the reference summaries. The results are shown in Table 1.

Table 1: ROUGE Scores for Summarization

Metric	Precision	Recall	F-Measure
ROUGE-1	0.2 (20%)	0.222 (22.2%)	0.211 (21.1%)
ROUGE-2	0.0	0.0	0.0
ROUGE-L	0.1 (10%)	0.111 (11.1%)	0.105 (10.5%)

5.3 Analysis

- **ROUGE-1:** The summaries capture some important words from the reference but have low overall alignment.
- **ROUGE-2:** The absence of bi-gram overlaps suggests structural or phrasing mismatches between the generated and reference summaries.
- **ROUGE-L:** Minimal alignment in sequence structure indicates room for improvement in capturing the reference's overall flow.

5.4 Discussion

- **Strengths:** The transcription model achieved a strong Word Error Rate (WER) of 12.4%, highlighting its robustness on clean audio data.
- **Weaknesses:** Summarization results showed limited alignment with reference summaries, especially for bi-grams (ROUGE-2) and sequence structure (ROUGE-L). Improvements are needed to enhance both factual retention and fluency.

6 Conclusion

This project implemented a speech-to-text transcription and summarization pipeline using state-of-the-art models. The Wav2Vec2 model achieved a Word Error Rate (WER) of 12.4% on clean audio, while the hybrid summarization approach combining extractive (BERT) and abstractive (BART) techniques achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 21.1%, 0.0%, and 10.5%, respectively.

6.1 Impact of the Project

The pipeline addresses challenges in transcription and summarization, offering a modular solution adaptable to real-world applications such as meeting transcription, virtual assistants, and accessibility tools.

6.2 Limitations

- Summarization performance, particularly in bi-gram overlap (ROUGE-2) and sequence structure (ROUGE-L), was low. This indicates a need for better alignment between generated and reference summaries.

6.3 Future Directions

- **Fine-Tuning:** Enhance summarization performance by fine-tuning the model on domain-specific datasets with high-quality reference summaries.
- **Hybrid Approach:** Explore improved hybrid strategies to balance extractive and abstractive strengths.

6.4 Final Remarks

The project highlights the effectiveness of integrating transcription and summarization into a single pipeline, and provides a foundation for future enhancements to improve versatility and applicability. In this project we only cover the process, due to the processing power constraint more advance techniques could not be applied. This project interests me because of my less focusing ability while listening to speeches. I will continue working on this project, keeping trying to improve it.

References

A Appendix

You may include other additional sections here.