Accurate Subtitle Generation in videos

TranscribeAI
Ruizi Wang (wan01492@umn.edu),
Manpreet Singh (sing1174@umn.edu)
Sharan Rajamanoharan (rajam033@umn.edu)

1 Introduction

In this project, we aim to enhance the accuracy of subtitles in videos, especially for long sentences and complex speech scenarios. Captions are essential for accessibility, comprehension, and engagement, yet existing systems often fail to meet the mark when faced with fast speech, strong accents, or multiple speakers. Our objective is to develop methods that align text more accurately with spoken content, making video consumption seamless and inclusive.

1.1 Motivation

Speech-to-Text (STT) systems often encounter challenges in generating accurate captions, particularly in scenarios involving rapid speech, diverse accent backgrounds, multi-speaker interactions. These challenges impact the readability and comprehension of video subtitles, especially in educational and entertainment contexts. This research aims to enhance subtitle generation and accuracy by developing advanced transcription techniques that improve the overall quality of captions across diverse platforms.

1.2 Objectives

This research aims to enhance video subtitle alignment and segmentation for better accessibility and viewing experiences. Key objectives include:

- Developing advanced transcription models for accurate captions in complex audio.
- Implementing text segmentation techniques with state-of-the-art language models.
- Creating a robust methodology to validate caption quality across content types.
- Enhancing caption generation without altering original video timestamps.

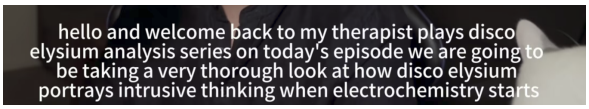


Figure 1: Text generated by STT model.

1.3 Previous Work

Speech-to-Text (STT) Systems

Whisper [1] was developed to address limitations of traditional automatic speech recognition (ASR) systems, particularly restricted dataset size and dependency on high-quality, human-validated transcripts. By leveraging 680,000 hours of labeled audio, Whisper enhances scalability and robustness of ASR models through weak supervision. This massive dataset includes diverse audio environments and speakers, allowing the model to generalize well across multiple languages and tasks, including transcription, translation, and voice activity detection, without requiring dataset-specific fine-tuning [1].

WhisperX [2] is an advanced STT system designed for long-form audio with accurate word-level timestamps. It incorporates Voice Activity Detection (VAD), parallel transcription and forced phoneme alignment to generate precise word timings. WhisperX improves transcription speed by twelvefold and supports multilingual transcription and alignment [2].

CrisperWhisper [3] improves upon Whisper's capabilities, particularly in timestamp prediction and tokenization. By employing Dynamic Time Warping (DTW), CrisperWhisper aligns audio segments with decoder token predictions more accurately. It also introduces a refined tokenization strategy, stripping spaces from tokens, which allows for better pause detection and facilitates more precise timestamp alignment[3].

Large Language Models (LLMs)

LLaMA[4] is a series of foundation language models developed by Meta AI. The models range from 7B to 65B parameters, with LLaMA-13B outperforming GPT-3 despite being significantly smaller. LLaMA 25 can handle larger contexts, with token limits up to 32,768 tokens for specific configurations [4] [5].

Claude is a family of language models developed by Anthropic, with architecture emphasizing safety and alignment, focusing on minimizing harmful or misleading outputs. Claude's architecture supports very high token limits, with Claude 2 offering up to 100,000 tokens.

Our work builds upon the advancements in STT models like Whisper and its variants to improve subtitle generation for videos. By leveraging the structured outputs (SRT/Json) from STT systems, our goal is to enhance caption segmentation and align them accurately with video frames, ensuring a seamless viewing experience.

1.4 Limitations

- Segmentation and Alignment: Difficulty in breaking sentences into readable segments often results in misaligned captions.
- Overlapping Speech: Simultaneous dialogue complicates voice differentiation.
- Background Noise: Environmental sounds interfere with speech clarity.
- Contextual Understanding: Limited grasp of context can result in misleading captions.
- Technical Constraints: Real-time processing and scalability across platforms are significant barriers.
- User Experience: Poor timing or inaccuracies reduce viewer engagement.
- Quality Consistency: Maintaining readability and accuracy across diverse speakers is difficult.

1.5 Potential Impact

This project aims to improve accessibility for deaf and hard-of-hearing viewers, enhance comprehension in educational contexts, and help content creators reach global audiences with precise English subtitles. Accurate captions can increase viewer engagement, boost SEO, and ensure a seamless user experience. Additionally, the project could advance AI and NLP technologies by refining captioning methodologies, benefiting video content quality across platforms.

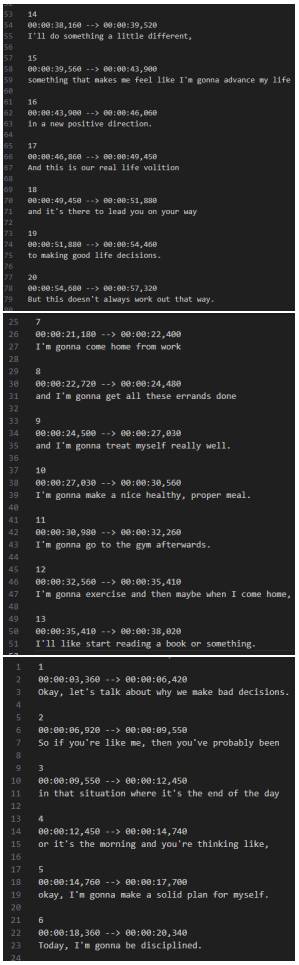


Figure 2: Examples of ground truth SRT files.

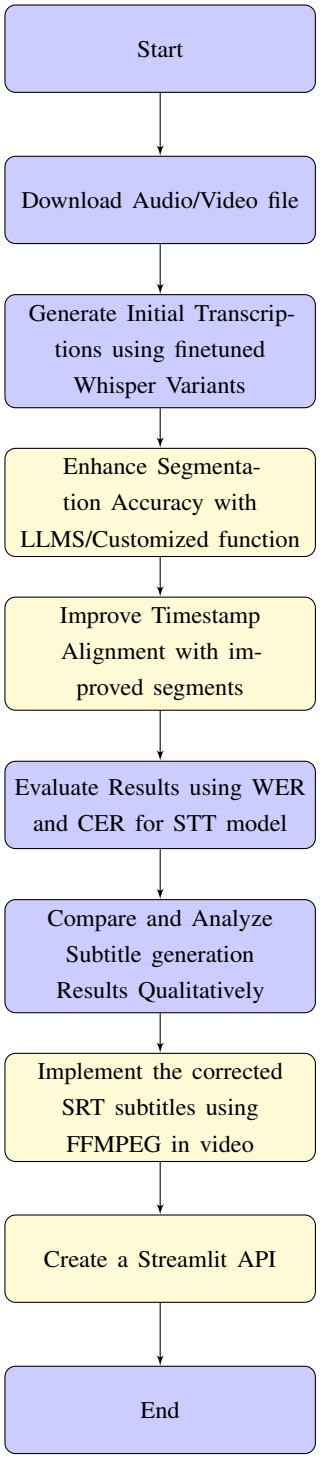
2 Novelty

Our project addresses limitations in Whisper's transcription process, focusing on improving spoken word accuracy and subtitle readability. Whisper, while robust, often struggles with punctuation, grammar, and contextual understanding, leading to less polished subtitles.

To enhance accuracy, we integrate Whisper with LLMs like Gemini-1.5 and GPT-40, for improving segmentation of long sentences. By experimenting with different prompting techniques we are able to be coherent captions with correct grammar and punctuation for relevant video frames.

We also introduce a custom segmentation function to split long speech segments into concise, contextually meaningful chunks, improving readability and comprehension. Our approach delivers

more accurate, user-friendly subtitles, addressing critical challenges in transcription and subtitling for media, education, and accessibility. Figure below represents our workflow diagram for our project:



3 Approach

3.1 Dataset and Pre-Processing

Custom Dataset:

We constructed a custom dataset from YouTube videos created by Euro Brady [10], which focus on analyzing the game *Disco Elysium*. The dataset comprises:

- **13 long videos** (1–2 hours each): Offering in-depth analysis of the game.
- **3 short videos** (20–30 minutes each): Providing concise insights.

Pre-Processing:

To prepare the dataset, we used automated transcription tools to generate `.srt` files, which were subsequently refined manually by:

- Aligning timestamps with spoken content for precise synchronization.
- Correcting errors in proper nouns, grammar, and punctuation to enhance transcription accuracy.

This process ensured a high-quality dataset for evaluating and improving subtitle generation techniques.

3.2 Methodology

In our approach, we compared the ground truth (GT) captions with transcription outputs from three different variants of the Whisper model:

- Whisper-X
- Whisper-Timestamped
- Whisper-Medium

To improve the performance of these models, we fine-tuned the Whisper base models on the **PolyAI/minds 14 dataset**, which consists of a large collection of conversational speech data from diverse sources. This dataset includes over 14,000 hours of dialogue across multiple domains and accents, making it an ideal resource for training a more robust speech-to-text (STT) system.

Fine-tuning on this dataset enabled Whisper to better handle various speech patterns, accents, and noise in real-world audio, leading to more accurate transcriptions, especially in noisy environments.

3.2.1 Whisper-Timestamped and VAD Feature

Among the three models, **Whisper-Timestamped** stands out due to its ability to provide precise timestamps for each word in the transcription. It also integrates **Voice Activity Detection (VAD)**, which helps the model distinguish between speech and silence.

The VAD feature is a significant improvement over the other two models, Whisper-X and Whisper-Medium, which do not have built-in VAD functionality. VAD helps process audio more efficiently, like detecting and skipping silent or non-verbal sections, which reduces errors and enhances transcription quality, especially in dynamic, real-time conversations. Thus we can get more accurate and cleaner transcriptions for generating subtitles.

3.3 Segmentation Approaches

For subtitle generation, we experimented with multiple segmentation approaches to handle long sentences and ensure proper alignment of subtitles.

Gemini 1.5 Pro[11]

We used the full transcript from the Whisper-Timestamped model as input for Gemini 1.5 Pro, instructing it to segment the text **without modifying the extracted content**. While Gemini is effective at generating accurate segments in its standard form, we encountered challenges with longer videos (over 5 minutes in duration), where it tended to produce hallucinations or repeat the original text without proper segmentation.

SaT[9] (Segment Any Text model)

Segment Any Text (SAT) is a fast, multilingual sentence segmentation model that handles noisy and short text efficiently. It uses subword tokenization, robust training with data corruption, and a limited lookahead mechanism for accurate segmentation across various domains.

We have finetuned SaT on our **Custom dataset** to observe its performance for subtitle segmentation and alignment.

Custom Segmentation Function

We developed a custom segmentation function that helps in generating improved subtitles without any LLMs. To improve readability and alignment we have incorporated the following into our function:

- **Splitting long segments:** Splits long speech captions into smaller, manageable parts for subtitle generation. Divides captions based on punctuation (e.g., commas, periods). Further splits phrases if they exceed a specified word limit (`max_words`).

- **Timestamp Synchronization and Optimization:** Distributes timestamps proportionally based on the word count in each segment.

- **Identifying pauses:** Inserts pauses between segments if a gap between them exceeds the specified `gap_threshold`. Creates new subtitle segments at detected pauses that exceed the threshold for better synchronization.

```
prompt = f"""
    You are a helpful assistant. Your task is to segment sentences which is
    longer than 65 characters, including spaces and punctuation, into shorter
    sentences. Each segmented sentence MUST NOT exceed 65 characters.
    Each segmented sentence must be independent, complete, and clear, suitable
    for direct translation or subtitle creation.
    NOTE: All connecting words (e.g., where, which, and, but, that) MUST remain
    intact. They should NOT be omitted, split, or modified in any way.

    Example:
    Input:
    so one of the biggest challenges when talking about anything related to
    mental health whether depression mental illness addiction is it can
    sometimes be really hard to explain how can you want something
    Output:
    so one of the biggest challenges when talking about
    anything related to mental health
    whether depression mental illness addiction
    is it can sometimes be really hard to explain
    how can you want something:{text}
    """
```

Figure 3: Prompt used with LLMs for segmentation: Combining Role-play and Contextual Fusion

3.4 Anticipated Challenges & Limitations

1. **SRT Input Issues:** LLMs like Gemini 1.5 Pro and GPT-40 altered timestamps when processing SRT files, causing misaligned and desynchronized subtitles.

2. **Processing long videos:** For videos over five minutes, LLMs struggled with segmentation accuracy, frequently hallucinating or repeating text input or given prompt.

3. **Unstable Output from LLMs:** Outputs from GPT-40 and Gemini were inconsistent, which necessitate the need for advanced prompting techniques.

4. **SaT Limitations:** Even after fine-tuning, SaT produced overly long segments unsuitable for subtitles, requiring extensive SRT data beyond our resources.

3.5 Scientific Novelty in Approach to Address Challenges

1. **Optimized LLM Usage:** Advanced prompting (role-playing, contextual fusion- prompt shown in Figure 3) and chunking via asyncio improved Gemini's segmentation stability and efficiency.

2. **Custom Segmentation Function:** We created a lightweight, non-LLM approach to split captions dynamically, optimize timestamps, and detected pauses for better subtitle readability and alignment.

3. **Fine-Tuning Insights:** Highlighted the limitations of pre-trained models like SaT in subtitle segmentation, validating lightweight custom solutions as resource-efficient alternatives.

3.6 Hypothesis

Our hypothesis is that combining custom segmentation techniques with fine-tuned STT models will outperform standalone STT models for subtitle generation. By refining segmentation based on punctuation, gaps, and timestamps, this hybrid approach aims to improve alignment and accuracy, especially for accents, proper nouns, and fast speech.

4 Experiments and Results

4.1 Evaluation Metrics

1. Semantic Accuracy: We assess the final SRTs obtained after applying LLM/custom function segmentation w.r.t. GT files using:

- **Character Error Rate:** CER value indicates the percentage of characters that were incorrectly predicted w.r.t ground truth. The lower the value, the better the performance of the ASR system with 0 being a perfect score.

$$CharErrorRate = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct characters,
- N is the number of characters in the reference (N=S+D+C).

Figure 4: Character Error Rate Calculation

- **Word Error Rate:** WER is a similar metric derived from the Levenshtein distance, working at the word level instead of the phoneme level. WER values range from 0 to 1, where 0 indicates that the compared pieces of text are exactly identical, and 1 indicates that they are completely different.

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

Where:
- S stands for substitutions,
- I stands for insertions,
- D stands for deletions,
- N is the number of words in the reference (that were actually said).

Figure 5: Word Error Rate Calculation

Figure 6 shows the maximum and minimum WER and CER values obtained while evaluating the SRTs obtained from four different methods: Initial Whisper-timestamped SRT, and then corrected SRTs using Gemini-1.5, SaT model and our Custom Segmentation function.

Method	Max WER	Max CER	Min WER	Min CER
Whisper-timestamped	0.9090909091	0.6952650749	0.1020408163	0.04347826087
Gemini-1.5	0.7795918367	0.6414237031	0.1331828442	0.08157227388
Segment-any-text model	0.1673469388	0.1036118205	0.0612244898	0.02650038971
Custom Function	0.3367697595	0.1392222756	0.106122449	0.04347826087

Figure 6: Maximum and Minimum WER and CER values obtained with different methods

4.1.1 Observations from WER/CER Table

1. Whisper-timestamped transcripts exhibit high minimum and maximum error rates at both the word and character levels, reflecting significant inaccuracies.

2. Using Gemini-1.5 for segmentation reduces WER and CER but retains higher error rates due to potential insertions, deletions, or modifications of words and characters in the original transcript.

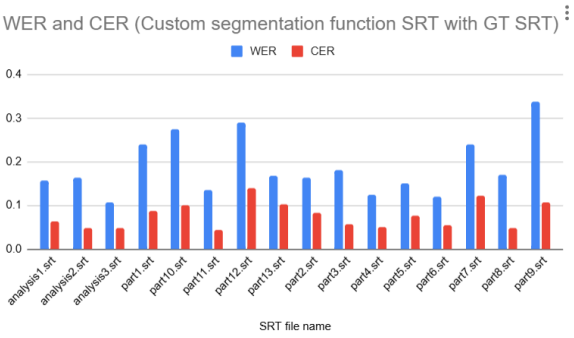


Figure 7: Visualizing WER/CER levels for different SRT outputs of Custom segment function

3. SaT model achieves the lowest error rates overall, though it tends to produce longer segments.

4. The Custom Function maintains low error rates with minimal modifications, primarily in punctuation, while ensuring segment lengths remain within the defined maximum word limit.

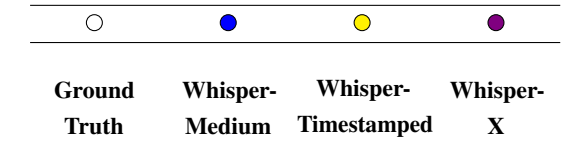
2. Performance of Different Segmentation Methods:

- **Segment Length Comparison:** Quantitatively comparing segment lengths generated by different methods (Whisper-timestamped, SaT, Gemini-1.5, Custom function) to assess readability. While segment alignment with the ground truth SRT was checked, varying segment counts across methods made direct matching difficult. Visual inspection on video was required to evaluate alignment effectively.

3. Human Evaluation: Human observation remains the most effective method to assess subtitle quality. This involves checking how well the subtitles align with the spoken words, ensuring the displayed text matches the speech, and verifying that subtitles appear on-screen in a timely manner, synchronized with the dialogue.

4.2 Results visualization

Displaying initial Whisper transcription results



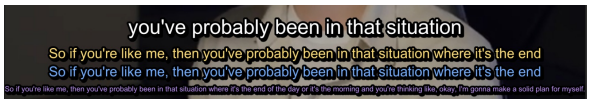


Figure 8: Initial transcription results - color coding represents the Whisper model used

Comparison of Segment Lengths in a Video

This illustrates the performance of different segmentation methods and their alignment with ground truth subtitles. Each method produces varying lengths of segments, thus the total number of segments also varies.

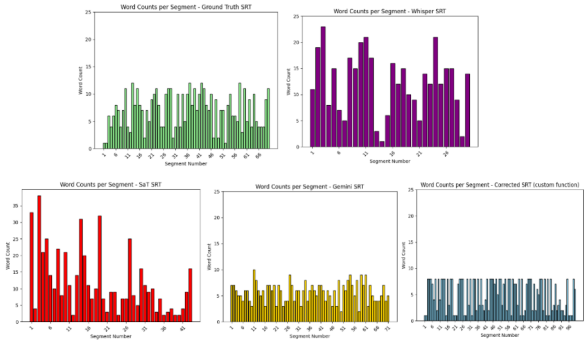


Figure 9: Comparing Segments across different Text Segmentation methods

Qualitative Analysis of Results:

Visually comparing outputs (example in Figures 10, 11, 12) revealed that the custom segmentation function closely matches Gemini's output, with only minor transcription differences in Gemini. However, like all LLMs, Gemini occasionally modifies the original transcript through word insertions or deletions.

The custom function demonstrated the most consistent performance, adhering to a maximum segment length of 8 words and maintaining sentence integrity and alignment.

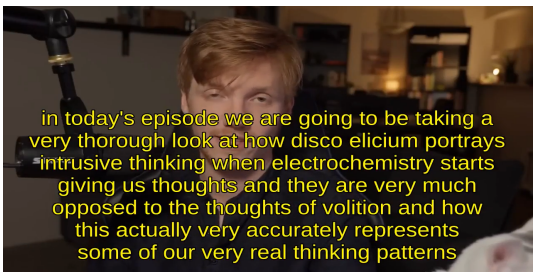


Figure 10: SaT output

Figure 11: Gemini-1.5 output

While Gemini and GPT perform comparably, SaT remains unsatisfactory despite being specifically trained for segmentation. It produces excessively long segments, making the subtitles unsuitable for viewing, highlighting the limitations of its current implementation for practical use.

Figure 12: Custom function

5 Error Analysis

Figure 13 illustrates a common issue with LLMs: the unintended insertion or deletion of words in the transcript.

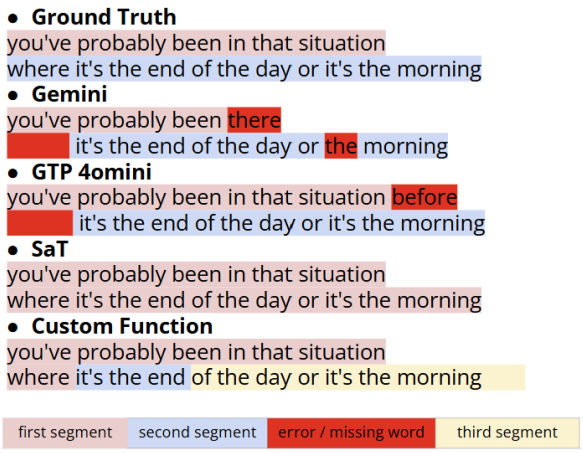


Figure 13: Results of each method for the same paragraph.

6 Discussion

6.1 Replicability

This work is easily replicable due to the detailed methodology and use of publicly available tools and datasets.

Dataset Contribution

The custom dataset[6], sourced from Euro Brady's YouTube series and enriched with manual refinements, includes accurate SRTs that serve as a robust baseline for benchmarking subtitle generation models. Its availability encourages further research into improving transcription accuracy across diverse audio-visual content.

Ethical Considerations

While the dataset comprises publicly available content, privacy and copyright laws must be respected. Inaccuracies in transcription, particularly in sensitive contexts, could lead to misinformation. Transparency about limitations and human verification for critical use cases are necessary to mitigate risks.

Streamlit GUI

We developed an end-to-end pipeline integrated into a user-friendly Streamlit GUI to demonstrate the Speech to text transcription and Subtitle generation process. Users can seamlessly input a YouTube URL or upload a video file, extract audio, and generate captions using Whisper-timestamped model and Segmentation function.

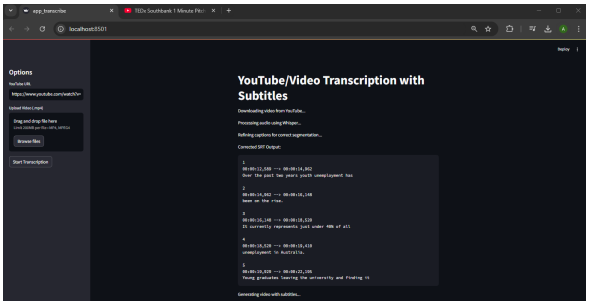


Figure 14: Display of GUI, downloading audio and video from a YouTube link given as input and generating Improved SRT file for Subtitles

Custom segmentation helps to refine the captions, ensuring accurate and well-structured subtitles. The refined SRT files are embedded into the video using FFmpeg, creating a captioned video output. With options to preview the video and download the SRT file, the pipeline offers a complete solution for showcasing the workflow and results in an interactive demo.

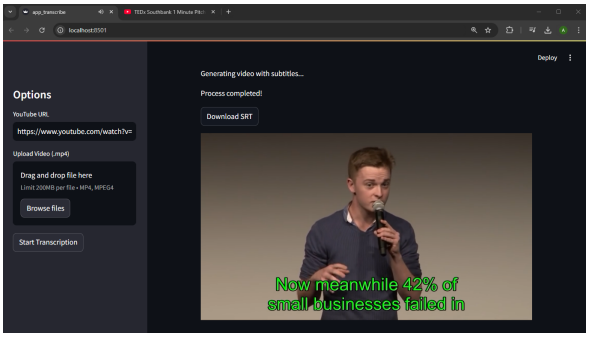


Figure 15: The final output video with added Subtitles

Future Work

Future research could focus on fine-tuning the Whisper/SaT models on diverse datasets for better accent recognition and handling complex audio environments. Implementing advanced noise-

canceling techniques and enhancing the contextual understanding of speech could further improve accuracy.

Also as suggested during the poster presentation, future work could involve ***comparing the audio signal with timestamps and transcribed subtitles to evaluate how accurately subtitles align with the speech flow***. However, this task is complex, as it requires simultaneously analyzing audio signals, word alignment, and timestamp accuracy. Due to time constraints, this approach could not be implemented in the current project but remains a valuable direction for further research.

7 Conclusion

7.1 Summary of Findings

This project improved subtitle accuracy by enhancing STT outputs with a custom segmentation function, achieving better word accuracy, punctuation, and readability. Metrics like WER and CER and comparison of segment lengths demonstrated that custom segmentation function and Gemini-1.5 have superior segmentation outputs compared to baseline models of Whisper and Segement any Text.

7.2 Significance of Results

Combining Whisper-timestamped with VAD feature transcription with multiple segmentation techniques (utilizing LLMs/ our custom segmentation function) proved effective in generating accurate, readable subtitles, addressing challenges like accents, fast speech, and noise. These findings contribute to advancing AI and NLP for automated transcription.

7.3 Potential Applications

The system has broad applications in media, education, accessibility, multilingual content creation, and corporate communication, enhancing transcription accuracy and user experience across diverse industries.

7.4 Github link

Here is the GitHub link to our project code and the zip file for the webpage: Subtitle_generation

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[2] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," *arXiv preprint arXiv:2303.00747*, 2023. [Online]. Available: https://arxiv.org/abs/2303.00747

[3] L. Wagner, B. Thallinger, and M. Zusag, "Crisper-Whisper: Accurate Timestamps on Verbatim Speech Transcriptions," in *INTERSPEECH 2024*, Aug. 2024, doi: 10.48550/arXiv.2408.16589.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. Singh Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] Brady, E. (Year), "Disco Elysium Video Analysis," YouTube Video Series, Available at: https://www.youtube.com/watch?v=177SU7ibNDg&list=PLBmriQSLAuRJvLNLhRBNso054qvpsqv3t

[7] M. Enis and M. Hopkins, "From LLM to NMT: Advancing Low-Resource Machine Translation with Claude," *arXiv preprint arXiv:2404.13813*, 2024.

[8] S. Wu, M. Koo, L. Blum, A. Black, L. Kao, F. Scalzo, and I. Kurtz, "A Comparative Study of Open-Source Large Language Models, GPT-4 and Claude 2: Multiple-Choice Test Taking in Nephrology," *arXiv preprint arXiv:2308.04709*, 2023.

[9] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, and M. Schedl, "Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation," *arXiv preprint arXiv:2406.16678*, 2024.

[10] Euro Brady, *Disco Elysium Analysis Playlist*, YouTube, 2024. Available at: https://www.youtube.com/watch?v=177SU7ibNDg&list=PLBmriQSLAuRJvLNLhRBNso054qvpsqv3t. Accessed: 2024-12-09.

[11] P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, S. Mariooryad, Y. Ding, X. Geng, F. Alcober, R. Frostig, M.

Omernick, L. Walker, C. Paduraru, C. Sorokin, A. Tacchetti, C. Gaffney, S. Daruki, O. Sercinoglu, Z. Gleicher, J. Love, P. Voigtlaender, R. Jain, G. Surita, K. Mohamed, R. Blevins, J. Ahn, T. Zhu, K. Kawintiranon, O. Firat, Y. Gu, Y. Zhang, M. Rahtz, M. Faruqui, N. Clay, J. Gilmer, J. Co-Reyes, I. Penchev, R. Zhu, N. Morioka, K. Hui, K. Haridasan, V. Campos, M. Mahdieh, M. Guo, S. Hassan, K. Kilgour, A. Vezer, H. Cheng, R. de Liedekerke, S. Goyal, P. Barham, D. Strouse, S. Noury, J. Adler, M. Sundararajan, S. Vikram, D. Lepikhin, M. Paganini, X. Garcia, F. Yang, D. Valter, M. Trebacz, K. Vodrahalli, C. Asawaroengchai, R. Ring, N. Kalb, L. Baldini Soares, S. Brahma, D. Steiner, T. Yu, F. Mentzer, A. He, L. Gonzalez, B. Xu, R. Lopez Kaufman, L. El Shafey, J. Oh, T. Hennigan, G. van den Driessche, S. Odoom, M. Lucic, B. Roelofs, S. Lall, A. Marathe, B. Chan, S. Ontanon, L. He, D. Teplyashin, J. Lai, P. Crone, B. Damoc, L. Ho, S. Riedel, K. Lenc, C. Yeh, and 1035 additional authors, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.