

Automatic Subtitle Generation from Movie Dialogues

Adrian Mincu and Huțanu Ciprian

Faculty of Mathematics and Computer Science

May 13, 2025

Overview

1. Abstract
2. Introduction
3. Approach
4. Conclusion
5. Future Work
6. Acknowledgements
7. Bibliography

We present a pipeline for automatic subtitle generation from movies, combining custom dataset creation with deep learning models.

Audio-dialogue pairs are extracted from full-length movies to build a domain-specific dataset. We infer DeepSpeech2 from scratch and fine-tune Wav2vec2 on this data.

Results show that domain-specific training and curated data significantly improve subtitle generation performance.

Manual subtitle creation is time-consuming and often inaccurate, impacting viewers who rely on subtitles.

Motivated by our own experiences as movie enthusiasts, we set out to automate subtitle generation to improve accessibility and enhance the viewing experience.

Our goal is to develop a faster, more reliable solution that benefits users seeking accurate, timely subtitles.

Recent advancements in Automatic Speech Recognition (ASR) have been driven by the development of large-scale, transformer-based models. Two notable contributions in this space are OpenAI's Whisper and Meta's Wav2Vec2:

- OpenAI's Whisper [Radford et al., 2022]
- Meta's Wav2Vec2 [Baevski et al., 2020]

Disclosure

The dataset used in this project is derived from Quentin Tarantino's films, which are known for their strong language, violence, and mature themes.

Viewer discretion is advised, as some content may be considered offensive or inappropriate. Please keep this in mind as you review the work.

Exploratory Data Analysis (EDA)



Figure: Word cloud visualization highlighting the most frequently occurring words in the subtitles text. Larger words represent higher frequency.

Exploratory Data Analysis (EDA)



Django Unchained (FFT)



Jackie Brown 1 (FFT)



Jackie Brown 2 (FFT)



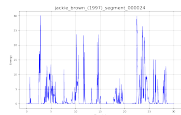
Pulp Fiction 1 (FFT)



Pulp Fiction 2 (FFT)



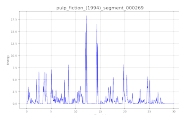
Django Unchained (Energy)



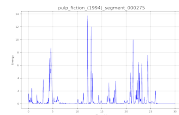
Jackie Brown 1 (Energy)



Jackie Brown 2 (Energy)



Pulp Fiction 1 (Energy)



Pulp Fiction 2 (Energy)

Figure: Top row – FFT analysis of audio signals; Bottom row – Corresponding energy distributions.

Data Pipeline

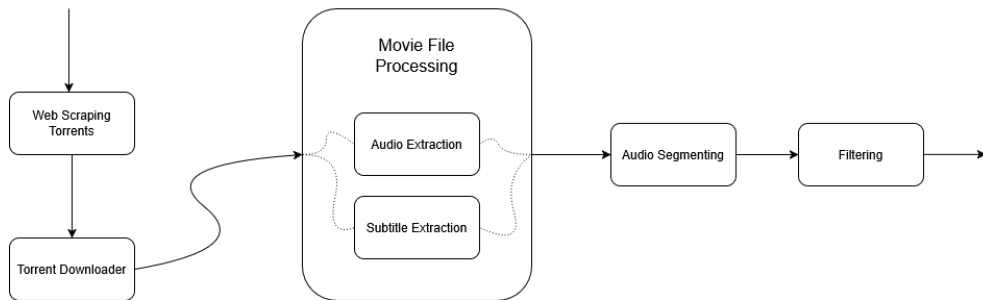
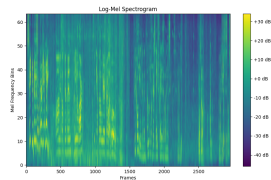
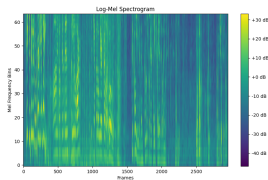


Figure: Overview of the data pipeline used for extracting, processing, and preparing movie audio for subtitle generation.

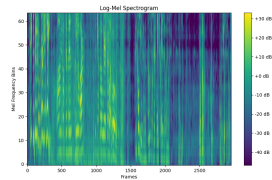
Audio Preprocessing



Raw Log-Mel Spectrogram



After Voice Frequency Filtering



After Voice Frequency Filtering + Wiener Filter

Figure: Visual comparisons of the spectrograms at various stages of preprocessing:

Subtitle Preprocessing

Before feeding subtitles into our model, we convert text into a numerical format. Traditional models like BERT use word-piece tokenizers with large vocabularies (tens of thousands of tokens), this creates a heavy computational burden for the model's output layer.

Instead, we use character-level tokenization with Wav2Vec2, which requires only 32 tokens—including the 26 English letters and a few special symbols like <pad>, <unk>, and | (used as a space).

This approach significantly reduces model complexity and is better suited for our speech-to-text pipeline.

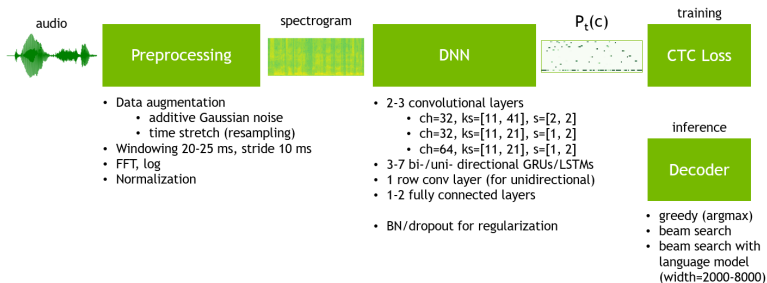


Figure: Overview of the Model Architecture for DeepSpeech2.

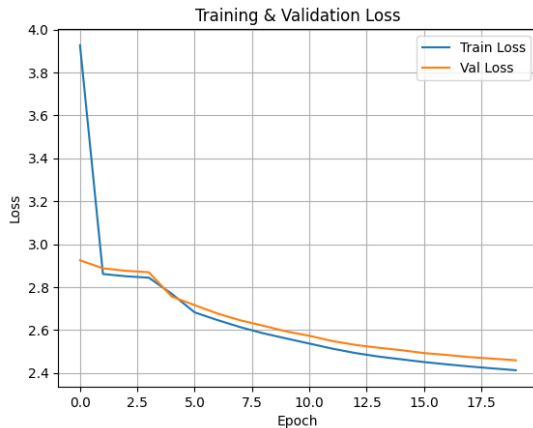


Figure: Training and validation loss curves for the DeepSpeech2 model during the subtitle generation process.

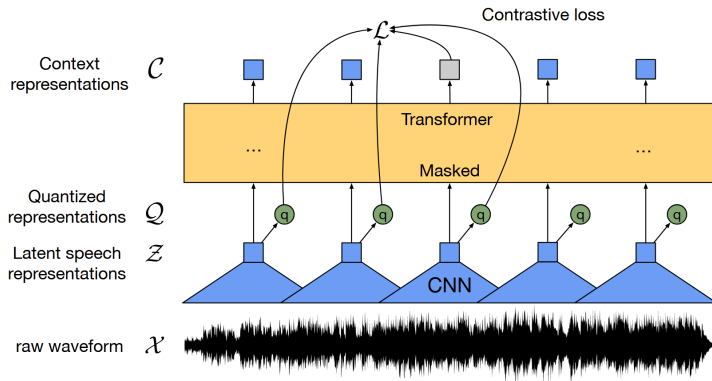


Figure: Overview of the Model Architecture for Wav2Vec2.

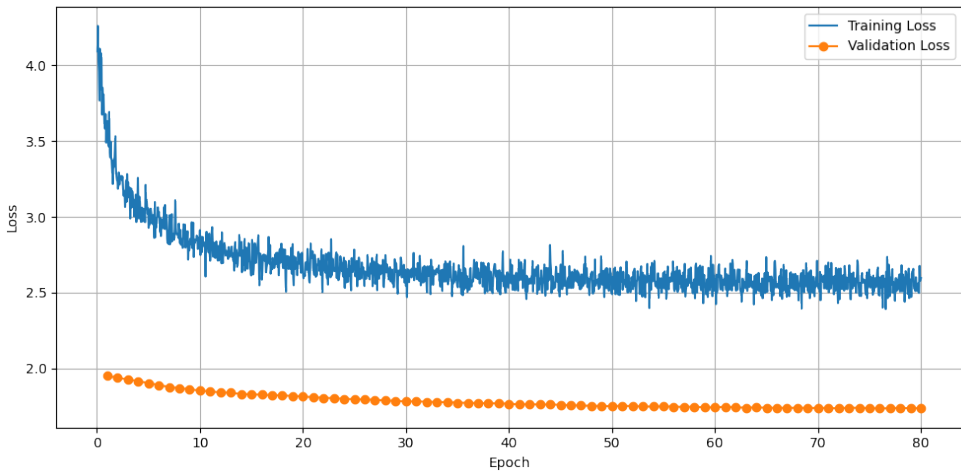


Figure: Training and validation loss curves for the Wav2Vec2 model during the subtitle generation process.

Comparative Analysis

Input: segment_000200.wav

Ground Truth:

*HOW'D YOUR LITTLE TALK WITH GEORGE GO? ARE WE KIDNAPPING HIM? NOT THE WORD I'D USE.
NOW YOU'VE TALKED TO HIM, YOU BELIEVE EVERYTHING'S ALL RIGHT? NOT EXACTLY. THIS WAS A
MISTAKE. YOU SHOULD LEAVE. WAY AHEAD OF YOU. GEORGE ISN'T BLIND. YOU'RE THE BLIND ONE.
GIRLS YELLING INDISTINCTLY.*

DeepSpeech2 Output:

*EA E E I E ORE A A E E E A E O EEA A O AE AE A AN AAN A O A A E E A E EA OEA I A ININ AN I E A SA
SA IN SOI IN I EI O E E E A E AIN AA ANAE*

Wav2Vec2 Output:

*HAV YOUR LITTLE TALK WITH GEORGE TO GO AR WE KID NAPPINGHAM NOT THE WORD I'D USE
WELL NOW THT YOU'VE TALKD TOHOME DO YOU BELIEVE EVERYTHING'S ALL RIGHT NOT EXACTLY
THIS WAS A MISTAKE YOU SHOULD LEAVE WHER GEORGE AND BLIND GER BLIND BLIND*

Observation:

- Wav2Vec2 output is significantly more intelligible and closer to the ground truth.
- DeepSpeech2 produces mostly unintelligible tokens.

Comparative Analysis

Segment	W2V2 WER	DS2 WER	W2V2 CER	DS2 CER
000200.wav	0.4285	1.3846	0.2867	0.8602
000184.wav	0.4000	1.0000	0.2134	0.8314
000047.wav	0.6626	1.2394	0.4644	0.8341
000096.wav	0.7215	1.0000	0.6066	0.8319
000001.wav	0.5588	1.1129	0.2849	0.8337
Average	0.5543	1.1474	0.3713	0.8383

Table: WER and CER results for Wav2Vec2 (W2V2) and DeepSpeech2 (DS2) on test segments.

Conclusion

This project taught us that building an end-to-end AI pipeline from scratch is no small feat, it's a bit like trying to edit a Tarantino movie: intense, full of challenges, and requiring a lot of patience.

But in the end, we learned that while AI pipelines can be tedious, it's all worth it, especially when you can binge-watch movies while working on them!

Future Work

While our results show promising transcription accuracy, future work could focus on aligning transcriptions with audio to generate time-stamped subtitles (e.g., .srt files).

This would enable:

- Automatic subtitle generation
- Real-time captioning
- Searchable movie dialogue databases

To achieve this, we could integrate tools like `aeneas`¹ for forced alignment or modify our models to predict timestamps directly.


¹Aeneas GitHub Repository


Acknowledgements


A special thanks to **Quentin Tarantino** for making such iconic films that not only redefine cinema but also provide us with endless hours of dialogue to transcribe. Without his remarkable ability to craft memorable lines, this project would have been much quieter, and much less entertaining.

Keep making movies, Quentin, we'll keep transcribing! [Tarantino, 1994]

Bibliography

 Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020).
wav2vec 2.0: A framework for self-supervised learning of speech representations.
arXiv.org.

 Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022).
Robust speech recognition via large-scale weak supervision.
openai.com.

 Tarantino, Q. (1994).
Pulp fiction.