AI Audio/Video Editor

Project Report
Presented to
The Faculty of the College of
Engineering

San Jose State University
In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Software Engineering

By
Kamtekar Yash
Ramesh Avinash
Shah Nevil

December 2022

Kamtekar Yash
Ramesh Avinash
Shah Nevil

APPROVED

Prof. Vijay Eranti

ABSTRACT

Collaborative AI Audio/Video Editor

by
Kamtekar Yash
Ramesh Avinash
Shah Nevil

AI is transforming the video production industry by making it quicker and easier to organize clips and create flawless edits. A Business Insider survey revealed that video production companies are changing: by 2018, 78% of marketers either used or planned to use AI [1]. Whether you're working on short-form movies or large-scale television shows, AI functionality will speed up the editing process and increase your creative options. Spend more time on the substance and less on editing. The good news is that artificial intelligence (AI) has developed techniques to expedite and lower the cost of production, such as automatic video editing, 3D animation, and realistic-looking visuals. Furthermore, it might make transcription easier and faster than it would be for a human to do it, saving money on labor costs.

Picture yourself recording a video and deciding to edit it before uploading it. There are a few ways to fix the problem. For example, you could record a completely new video or just the section you want to change, then combine it with the original. We frequently use filler words like "aa" and "..uhmm" when filming interviews or self-promotional videos, and we want to get rid of these when uploading the content or sharing it with others. There are times when we forget to record a part of a video and want to add specific speech or audio to the video; in these cases, we must rerecord the part and then endure the arduous process of editing the video.

In this project, we propose an all-in-one collaborative AI audio and video editing tool that is as user-friendly as editing a Google Doc and is based on text extracted from transcription as a solution to the aforementioned problem statement. With this method, AI can be edited by novice users with no editing experience like a pro. With the voice cloning for text-to-speech conversion, editing your recorded audio is as simple as

typing, and our suggested AI tool would have fantastic features like automatic transcription from audio, the ability to remove filler words with a single click, the addition of speaker labels, and text-to-speech conversion. Our solution entails identifying and comprehending all available cutting-edge state-of-the-art (SOTA) models, such as Google DeepMind's Flamingo multimodel **[2]**, attempting transfer learning for real-time voice cloning from existing text-speech models in research **[3]**, and developing an end-to-end ML product for our solution domain.

Acknowledgements

The authors are deeply indebted to Professor Vijay Eranti for his invaluable comments and assistance in the preparation of this study

TABLE OF CONTENT

List of Figures

Chapter 1. Project Overview

Introduction

According to a poll, 96% of marketers anticipate allocating additional funding for this channel, while 87% of marketers report seeing favorable ROI from videos. In addition, 69% of the experts polled who do not currently indulge in videos said they would soon. As a result, making movies will take more time, and getting the licenses for the tools that can make videos will cost more money.

People compelled to remain in their homes had to engage themselves, and 96% indicated a rise in the time spent watching movies. Brands immediately adapted and produced more content, mainly as AI video editing simplified it. However, this also means more rivalry, which pushes marketers to raise the bar for the caliber of their videos. The winning apps will be the ones that assist them in attaining it.

The practice of manipulating a video clip with machine learning and artificial intelligence is called "AI video editing." This may involve incorporating multiple color filters, augmented reality masks, and specific other enhancements and intelligently editing the video clip.

In this project, we offer a cloud-based application and a method for automatically modifying videos based on keywords gleaned from voice transcription. Video sequences are chosen and chained using an audio transcript to autonomously generate a fresh clip at a time specified by the user. According to the project's timing and viability, we intend to develop an all-in-one music and video editing tool that is as simple as editing a Google document.

Proposed Areas of Study and Academic Contribution

In this project, we propose a method for automatically editing videos based on keywords derived from voice transcription and a cloud-based application. A new clip with a time set by the user is automatically generated by selecting and chaining several video segments using an audio transcript. We plan to create an all-in-one music and video editing tool that is as straightforward as editing a Google document, depending on the project's timeframe and practicality.

We intend to use OpenAI Whisper, a general-purpose voice recognition model, for the Transcription(Speech-Text). It is a multi-task model that can do multilingual voice recognition, speech translation, and language identification and was trained on a sizable dataset of varied sounds.

Once we have the transcripts, we will use a custom model to stabilize the word timestamp. Most of our editing features, such as filler word removal, cropping, sub-clipping, and text-speech, are built on top of the word timestamps.

We intend to use transfer learning from speaker verification to Multispeaker Text-To-Speech Synthesis (SV2TTS) with a real-time vocoder to perform speech-text. SV2TTS is a three-stage deep learning framework that allows a numerical representation of a voice to be created from a few seconds of audio and used to condition a text-to-speech model trained to generalize to new voices.

We are incorporating speaker identification into our application. (Identifying speaker identity aids in answering the question, "Who spoke when?")

With the recent applications and advancements in deep learning over the last few years, it is now possible to automatically (and confidently) verify and identify speakers. Speaker diarization, combined with cutting-edge accuracy, can add enormous value to any mono-channel recording.

Speaker verification and dimerization were previously accomplished using vector-based audio embedding techniques.

However, with recent advances in deep learning, neural network-based audio embeddings (also known as "d-vectors") have proven to be the most effective method.

Specifically, LSTM-based d-vector audio embeddings with nonparametric clustering aid in developing a cutting-edge speaker diffraction system.

Current State of the Art

According to recent trends, the video production and consumption industries have increased drastically. Almost every market has used or plans to use AI in the future. Whether you are working on short videos or long films, AI capabilities will ease video editing and expand your creative options. Spend more effort on the content and less time on editing.

State-of-the-art in the machine learning/deep learning eras changes daily, with recent paper trends on various neural network capabilities. It has also enabled faster and easier transcription and translation services than humans, thereby speeding up production.

There is currently much-advanced research in the video consumption space for text-video generative models. Organizations such as Google Deepmind, Facebook research, IBM, Nvidia, Microsoft, and OpenAI were able to create powerful multi-models trained on billions of parameters using meta-learning and multi-task learning, graph neural networks, thanks to advanced research in AI space. We can see visually what the power of this model's DALL-E and OpenAI codex are capable of and where the AI is heading.

So, In this project, we propose a collaborative AI video and audio editing application based on text extracted from the transcription that is as simple to use as editing a Google Doc. With this approach, perhaps beginners with no editing skills can edit with AI like a champ. Instantaneous transcription from playback, the ability to eliminate filler words with the press of a button, voice cloning for text-to-speech conversion, which makes editing your recordings as simple as having to type, and the inclusion of speaker labels are just a few of the fantastic features that our proposed AI tool would include.

Numerous cutting-edge models exist in Speech-Text, Text-Speech, NLP, and translation tasks. Identifying and comprehending all available cutting-edge state-of-the-art (SOTA) models, attempting transfer learning for real-time voice cloning from existing text-speech models in research, and developing an end-to-end ML product for our solution domain are all part of our solution. Due to advancements in the field of transfer learning, we intend to use existing pre-trained models while modifying and adding a few existing layers to create SOTA multi-models for our problem domain.

Chapter 2. MLOps Architecture

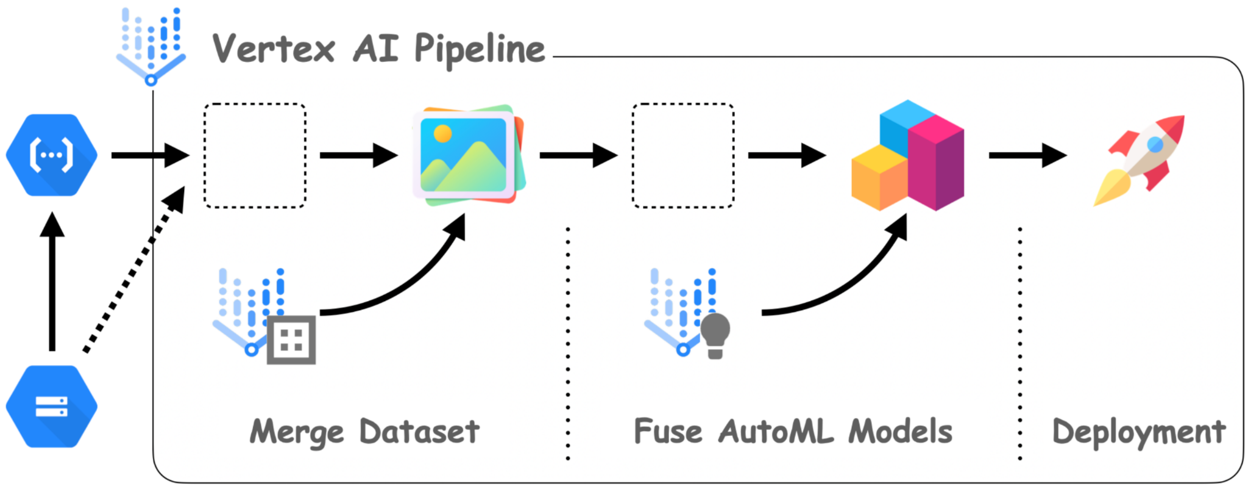

Figure 1: Simple MLOps architecture (Source)

Here, we studied several MLOps architecture for various different projects and performed literature survey for the same. Thereafter, we came across the MLOps architecture given by Vertex.AI ,which we have tried to follow during the entire duration of the project.

For the proof of concepts, we used many datasets like **LibreSpeech dataset** to perform Proof of Concepts (POCs) for various different features like speech-to-text transcription, anad attaching word timestamps. We have tested several pre-trained models like **OpenAI whisper** and **Descrypt APIs** for various services and found that OpenAI whisper gives the best accuracy. As we are using pretrained models, we do need to retrain the model continuously. Thereafter, we have used **streamlit** application for front-end development. Lastly, we created a **Docker Image** for the source code and thereafter, deployed the entire application on **HuggingFace**, so that any user can access the project and perform various operations in their audio/video input.

Chapter 3. Project Features

The Audio/Video software, which is deployed on HuggingFace and available for use to general public, has many different kinds of features. They are:

1) **Speech to Text transcription:** Transcription is converting audio/video recorded speech to text. Transcription, or transcribing as it is often referred to, is the process of converting speech from an audio or video recording into text. Transcription entails more than just listening to recordings. After the transcript is generated, we are also stabilizing the word timestamps for various other features. After performing proof of concepts, we have used OpenAI whisper model for generating audio/video transcripts
2) **Speaker Diarization:** Speaker diarization is a combination of speaker segmentation and speaker clustering. The first aims at finding speaker change points in an audio stream. The second aims at grouping together speech segments on the basis of speaker characteristics.
3) **Removing filler words from the video:** The first step is to upload video to generate the transcript, then get the audio from the video and generate transcript from the transcription service. Thereafter, attach the timestamp to the transcripts to remove filler words and join the audio and remove filler words and join the audio file.
4) **PII Redaction:** In this feature, the portion of the video which contains sensitive personal information is removed as per the user's request of type of information, which can be name, occupation, email address or phone number. Here, if the video contains any sensitive perforation information, which user doesn't want , then, he can select the input information type and the information will be redacted in the new generated video.
5) **Content Analyser:** With Summarization, we can generate a single abstractive summary of entire audio files submitted for transcription.With Topic Detection, we can label the topics that are spoken in your audio/video files. The predicted topic labels follow the standardized IAB Taxonomy, which makes them suitable for Contextual Targeting use cases. This API can predict the topic names among

698 different topics. With Content Safety Detection, we can detect if any of the following sensitive content is spoken in your audio/video files, and pinpoint exactly when and what was spoken.

Chapter 4. Project Requirements

Interface Requirements

User interface

An intuitive and collaborative AI audio/video editing online application that supports all features per document and was made using preferred frontend frameworks such that it can work on all current browsers without any issues.

Hardware Interface

Various cloud-computing services are required to host the application on the web.

Communication Interface

Various communications requirements between frontend and backend services for multiple AI features/services

Application Features

User Registration & Login

Users can continue as guests or there will be a sign-up page where users can create an account. During registration, the user must provide some information. Registration will be completed after validation, and the user will be notified.

User Authentication, Authorization and Session Management

There will be a login page where the user can enter his login information and access the system. Access the identity of a user, and determine their access rights for our application.

Dashboard

Various Activities carried out by the user in the application

Upload Media Files (audios/videos) to our application

The user would access our web app and upload his audio/video files.

Support various Media Formats

The application should be able to handle below common media formats MP4, MOV, WMV ,AVI, AVCHD, MKV, WEBM or HTML5

Life Cycle Management for the Media files

Setup Life cycle Management for the Uploaded Videos based on the user access rights.

Editing Video like pro with the help of AI

Edit videos with transcription from audio/video

Edit Panel would show the transcription of the audio/video. Each transcription is linked to the audio/video content's timeline. If he wants to trim the contents, he simply removes the transcribed text contents from the video, and the video is automatically trimmed.

Remove Filler Words from video

Users can remove the common filler words from the video with a click of a button and all the contents are reedited and trimmed automatically.

Identify and add Speaker Labels

Make it easy to automatically or manually add speaker labels to your transcripts.

Replace Existing content with new Contents

If a user wants to replace existing contents with new contents, he can select the transcription at a specific timeline and replace existing contents with new contents with a single click, eliminating the need to record new audio or video.

Additional features

There are also plans to include features such as Entity identification, PII redaction, video enhancement, and so on which we would work on later parts of the course.

Performance requirements

The application's performance should be optimized. Our expectations should be met or exceeded in terms of application response time, throughput, execution time, and storage capacity. Because the application must be web-based and run from a frontend/backend application service.

The app's initial load time will be determined by the strength of the internet connection, which is also determined by the media from which the app is run. The performance will be determined by the client's/hardware customer's components.

Non Functional Requirements

Software system attributes

Security and Privacy

Because each user has an account that stores personal information and activities, the system maintains a high level of security against unauthorized access, authorization, and authentication. Furthermore, the system website may include an SSL certificate, allowing the website to run securely on the https protocol.

Efficiency and Usability

This system ensures high efficiency, which was achieved through extensive testing and iteration of the user experience flow. By reducing the number of clicks required to reach the end goal and displaying each stage on a visually intuitive interface, a satisfactory goal was achieved.

Scalability and Performance

Scalability was considered when designing the system. It should be able to handle hundreds of thousands of concurrent connections without issue. All Cloud Services facilitate scalability based on our requirements and costs.

Chapter 5. Dependencies and Deliverables

Dependencies

- Data Collection and Feature Extraction
- High Computation
- Model Reliability
- Cost Management
- Cleaning and preprocessing the data and convert it into a required format
- The cold-start problem

Deliverables

- Automatic transcription from audio and video
- Ability to remove filler words with the click of a button
- Text-to-speech conversion with custom voice cloning
- Delivering more improvised User Interface and User Experience Interface
- Additional Features like Speaker Labels identification, PII redaction.

Chapter 6. Project Design

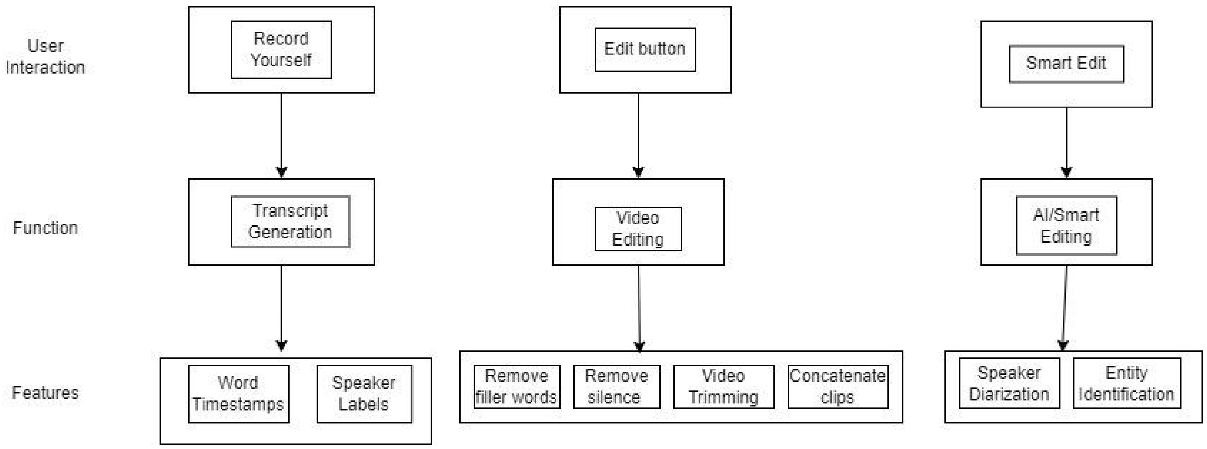

Figure 2: Project Design

The system design consists of 3 layers:

1) **User Interaction**: The user interaction consists of users recording the audio/video, can edit the video, or can utilize smart editing options.

2) **Function**: The function layer intends to determine what is the motive of the input given by the user. The functions here might include video editing features , transcript generation or smart editing.

3) **Features**: The feature layer contains the total features which the video editing software is capable of performing on the basis of input given by the client. The features can include adding the speaker labels, video trimming, adding the timestamp to each word or speaker diarization.

Chapter 7. Individual Contribution

1) **Avinash Ramesh:** I started working on the different AI services like content Analysis and Personally Identifiable Information (PII) redactation.For all those features, I tested different state of the art models for these features and finalized the pretrained model as per its performance and accuracy. I also performed the worked on the Proof of Concepts (POCs) of integrating all AI services and testing out all the functionality together and see how it works using a simple streamline application.

2) **Nevil Shah:**I was entrusted with the responsibility to generate audio/video transcripts from an audio/video and try various state of the art models for audio/video transcription and determine its accuracy to the original content. First and Foremost,I performed the proof of concept for various models like openAI whisper and descrypt APIs and found that openAI whisper model worked the best. Firstly,after saving the live recorded input file, the transcript for that input is generated.I also tried the same process for different types of audio with regards to speed of speech, noise variation and the tones. Thereafter,I attached timestamps to every word in the generated transcript.Lastly, I added speaker labels to each of the different voices in the audio file.

3) **Yash Kamtekar**: I developed the PoC of filler word removal service on click of a single button. This feature works audio/video file that is either recorded or a YouTube video. It extracts the audio from the file and then generates the transcripts after which it attaches timestamp to these transcripts for further processing. The generated output from the transcription service is given to the Open Whisper model which has been given a small set of filler words. The model prints the filler word and it’s duration as per the audio generated from step 2. Using moviepy library of python and the output of the model the audio files are merged to generate the output without the filler words.

References

[1] The Future of Video Production: AI, Big Data & Machine Learning

[2] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

[3] Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in neural information processing systems*, *31*.

[4] Q. Xie et al., “The Multi-Speaker Multi-Style Voice Cloning Challenge 2021,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021, doi: 10.1109/icassp39728.2021.9414001.

[5] H.-T. Luong and J. Yamagishi, “NAUTILUS: A Versatile Voice Cloning System,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2967–2981, 2020, doi: 10.1109/taslp.2020.3034994.

[6] C. Bokhove and Christopher, “Automated generation of ‘good enough’ transcripts as a first step to transcription of audio-recorded data,” Jan. 2018, doi: 10.31219/osf.io/sn7w9.

[7] V. Morfi and D. Stowell, “Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets,” Applied Sciences, vol. 8, no. 8, p. 1397, Aug. 2018, doi: 10.3390/app8081397.

[8] Tanberk, S., Dağlı, V., & Gürkan, M. K. (2021, September). Deep Learning for Video Conferencing: A Brief Examination of Speech to Text and Speech Synthesis. In *2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 506-511). IEEE

[9] Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

[10] Chen, Y., Assael, Y., Shillingford, B., Budden, D., Reed, S., Zen, H., ... & de Freitas, N. (2018). Sample efficient adaptive text-to-speech. *arXiv preprint arXiv:1809.10460*.