Automated Video Segmentation

A TECHNICAL REPORT

submitted by

Arjun Sadashiv - AM.EN.U4EAC20012
Arun Gangadharan M - AM.EN.U4EAC20013
Jibin Thomas Daniel - AM.EN.U4EAC20030
Sanat Jayakrishnan - AM.EN.U4EAC20058

under the guidance of

Ms. Anuraj K and Dr. Vivek Venugopal

submitted as part of

19ECE495/ 19EAC495 PROJECT PHASE I

in

ELECTRONICS AND COMMUNICATION ENGINEERING

AMRITA SCHOOL OF ENGINEERING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI (INDIA)
November - 2023

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

AMRITAPURI (INDIA)

BONAFIDE CERTIFICATE

This is to certify that the report entitled **"Automated Video Segmentation"** submitted by **Arjun Sadashiv(AM.EN.U4EAC20012), Arun Gangadharan M(AM.EN.U4EAC20013), Jibin Thomas Daniel(AM.EN.U4EAC20030), Sanat Jayakrishnan(AM.EN.U4EAC20058)** as part of the 19ECE495/ 19EAC495 PROJECT PHASE I is a bonafide record of the work carried out by her under my guidance and supervision at Amrita School of Engineering, Amritapuri.

Signature of Supervisor:
Ms. Anuraj K Assistant Professor Sr. Grade
Amrita Vishwa Vidyapeetham, Amritapuri
Department of ECE

Signature of Co-Supervisor:
Dr. Vivek Venugopal
Assistant Professor
Amrita Vishwa Vidyapeetham, Bengaluru
Department of ECE

Date: 27/11/2023

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

AMRITAPURI - 690 542

DEPARTMENT OF ECE

DECLARATION

We, **Arjun Sadashiv(AM.EN.U4EAC20012), Arun Gangadharan M(AM.EN.U4EAC20013), Jibin Thomas Daniel(AM.EN.U4EAC20030), Sanat Jayakrishnan(AM.EN.U4EAC20058)** hereby declare that this technical report entitled **"Automated Video Segmentation",** is the record of the original work done by us under the guidance of **Ms. Anuraj K,** Assistant Professor Sr. Grade, Department of ECE, Amrita School of Engineering, Amritapuri and **Dr. Vivek Venugopal,** Assistant Professor, Department of ECE, Amrita School of Engineering, Bengaluru. To the best of my knowledge this work has not formed the basis for the award of any degree/diploma/ associateship/fellowship/or a similar award to any candidate in any University.

Place:Amritapuri
Date: 27/11/2023

Signature of the Students

Acknowledgement

During the creation of our model, we faced our own trials and tribulations. Each of which was overcome by the immense support, wisdom and advice we received from our guides, **Dr. Vivek Venugopal and Ms. Anuraj K**. Their expertise on the subject matter was irreplaceable and their encouragement pushed us to overcome our limitations, words cannot describe the gratitude we would like to extend towards our guides. We would also like to extend heartfelt to the many individual contributors whose work served as stepping stones for us to base our work off of. We would also like to thank our families for their unwavering support and for understanding the demanding nature of our project. Their belief in us was our guiding star that kept pushing us forward.

We would also like to extend our respect and appreciation to the **Amrita School of Engineering**, for arming us with the knowledge and wisdom necessary to complete our project. The incredible learning environment of our university and the resources provided to us by it played a crucial role in shaping us in preparation for this project. We would also like to express our gratitude to our dear friends and peers whose inquisitiveness and interest in our project pushed us to do our best. Last but not least we would like to thank all the authors and contributors whose work in the field of Video Segmentation using Deep Learning laid the foundation for our work. Their work has been essential for the completion of our work. Our Automated Video Segmentation using Audio Cues was made possible only due to the immense support we received and because of the coming together of all of these various factors. Finally, we would like to extend our unwavering gratitude to anyone and everyone that has helped along our path to completing our project.

Contents

Abstract

Online lecture videos have become an overwhelming juggernaut in the online content creation spheres of popular video streaming sites. Especially after the recent Covid-19 pandemic that shifted the physical classroom dynamic to that of a virtual one. Without the organic interaction between teacher and student that would have occurred in a physical classroom setting, many a learning experience was stifled, and information could not be conveyed efficiently. This led to a large number of students relying on the Internet to learn and a large number of them flocked to the video lectures that were available on streaming platforms like MITOCW, Coursera, NPTEL, and YouTube to name a few. Many of these lecture videos are too lengthy and might cover multiple concepts. So, it is imperative to segment the videos for increased adoption among the student community. The main objective of our project is to create an automatic applied lecture video segmentation system that splits lectures into bite-sized topics using audio cues. Prior works in this regard have considered the video frames and their output from an Optical Character Recognition (OCR) system for segmentation. To the best of our knowledge, no work has been explored using the audio from the video lectures for this task. The performance of our proposed system will be computed using metrics such as Normalized Mutual Information (NMI), Mean over Frames (MoF), F1-score, and Intersection over union (IoU).

Chapter 1

Introduction

1.1 Introduction

In a regular household, with the average family, education of the children tends to be of paramount importance. In today's day and age, where electronics and the internet run rampant in every household, education and access to information have become incredibly easy. One such method of extracting information is by simply watching educational. videos in which a lecturer explains a concept and understanding it. This has become the lifeline of many a student in the far-reaches of the world. But in today's busy world, ease of use and convenience also tend to play a major role in the way students tend to curate the video lectures they require or select to parse through. Convenience usually translates how quickly one can go through one such video and the segmentation of a lecture usually helps a student in increasing their ease of use. But segmentation of video lectures can only occur if the streaming platform the video is posted on permits and if the creator of the video does segment their lecture. Now this is not the case most of the time, video lectures do not come pre-segmented. This is an inconvenience we aim to resolve using our project. By understanding and reviewing similar works done in the same field we seek to create a model which automatically segments a video based on the topics covered within the lecture and its correlating audio. In this review, we highlight some papers and summarise the contributions made by them as they all will help us in our journey to build an audio-analysing model.

Chapter 2

Literature Survey

2.1 Topic Transition In Educational Videos Using Visually Salient Words

[1] The rapid growth of online courses and Open Educational Resources (OER) has led to a need for methods to efficiently consume this multimedia content. One of the most challenging areas of research is to automatically identify topic transitions in educational videos. Topic transitions are the points in a video where the topic changes. For example, in a lecture video on types of discrete random variables, the topics can vary from Binomial RV, Geometric RV to name a few. Manually identifying topic transitions is a time-consuming and expensive process. A new approach to automatically identify topic transitions in educational videos is proposed in this paper. The approach analyses the visual content of a lecture video to determine the transition points.

2.2 Localizing Moments In Video With Natural Language

[2] The paper discusses on localizing or segmenting certain moments in a full-length video. The aim is to retrieve or capture a particular segment from a video using natural language textual inputs by the user. For this they proposed a model known as Moment Context Network (MCN). MCN model localizes the input queries in the video by integrating local and global features from the video over time. The moment retrieval

algorithm localizes the queries in longer videos. The visual temporal context features are extracted which encode the moment in the video by integrating the local and the global features of the video. The language features are extracted using LSTM techniques. Visual Temporal Context Features: the video moments are encoded into visual temporal context features by integrating local video features, global video features and temporal endpoint features. The language features are extracted using LSTM techniques. After the encoding the LSTM hidden state is passed through a single connected layer to yield embedded features, which is passed to a Joint Video and Language Model. The joint model is the sum of the squared distance between the embedded appearance, flow and language features. The model is trained with a ranking loss which compares the closeness of the referring expressions with its corresponding moments from the video than negative moments in a shared embedding space.

2.3 Automatic Hierarchical Table Of Contents Generation For Educational Videos

[3] The paper proposes a new method for automatically generating a hierarchical table of contents (ToC) for educational videos. The ToC is a summary of the video content, organized in a hierarchical structure, that allows users to quickly navigate to the desired section of the video. The method consists of mainly three steps. Shot segmentation: In the first step the whole video is segmented into short consecutive frames which are visually familiar called shots. Text extraction: The second step involves extraction of textual information from each shot, like the speaker's text, slide titles, captions. ToC generation: The third step is to generate a hierarchical ToCusing the information from each shot. A novel method for shot boundary detection is proposed. The method consists of three steps: visual content representation, continuous signal construction, and classification of shot boundaries. In the visual content representation step, the shape of objects in an image is captured by computing the entropy of the edge map in a patch-wise manner. Color variations are captured by representing the image with the

2D histogram of hue and saturation values. In the continuous signal construction step, a measure is computed to capture the resemblance between two frames. This measure is a product of the similarity metrics for color and shape features. In the classification of shot boundaries step, a clustering-based technique is used to choose an adaptive threshold for classifying whether a boundary is an shot boundary or not. Visual text extraction is performed using a commercial OCR engine and a deep learning model to generate a heat map of text regions in an image. The text regions are then filtered based on their overlap with the bounding boxes from the OCR engine. The salient features of each visual text entity (vText) are extracted, including font size, boldness, and vertical location. All the vTexts in a shot are aggregated into a single set to avoid processing repeated texts. The final ToC is generated by combining the visual and audio text information.

2.4 End-To-End Learning Of Visual Representations From Uncurated Instructional Videos

[4] In this paper the authors propose a new learning approach, MIL-NCE, which can deal with misalignments in narrated videos. This is aimed to make strong video representations, without having to do any manual annotation. A 3D CNN backbone is used here, specifically the I3D implementation. The visual representations are used at two different levels, the first being the output of the I3D Global avg pool and the second being the output of the final I3D layer. The visual representations are used at two different levels, the first being the output of the I3D Global avg pool and the second being the output of the final I3D layer. To depict the all-roundedness of the representations, an evaluation is performed on five downstream tasks: Action Recognition, Text-to-video Retrieval, Action Localization, Action Step Localization, Action Segmentation.

2.5 Unsupervised Audio-Visual Lecture Segmentation

[5] The paper proposes a video lecture segmentation model that splits a lecture into bite-sized components. The clip representations use visual and OCR cues and the model is trained on a pretext self- supervised task of matching the narration with temporally aligned visual content. TW- FINCH, a variation of the 1- Nearest Neighbour algorithms is a utilized on these representations. The unique position this model occupies in the sphere of research relies solely on the unique model that has been developed and trained for specific purpose. The approach highlighted in this paper has been split up into three different stages. The three key feature types used here are OCR, 2D and 3D. The OCR feature encodes the output text from an OCR API using the BERT sentence transformer model. The Text Feature Extraction The model is built by first extract the visual and textual features for a video clip C and transcript (text) T using the feature extraction pipelines described above. The OCR feature is then passed through a fully-connected layer to obtain a 2048- dimensional vector. The model's parameters is trained train with the max-margin ranking loss function. From the joint text-video model the clip is extracted and is transcripted. The representation of a lecture with N clips are transferred to the TW-Finch algorithm which is then represented as a 1-Nearest-Neighbour graph after encoding of feature similarities and temporal proximities. In training, the first stage encapsulates the pre-training of the model using the CwoS dataset. The second stage includes pre-training with the CwS dataset in an unsupervised manner. In evaluation dataset all 15 courses of CwS to understand and derive the performance of the model. To measure the overall accuracy of the model metrics such as NMI (Normalized Mutual Information), MoF (Mean Over Frames), F1-score, IoU (Intersection Over Union). The result of this Unsupervised Lecture Segmentation is the visualization of segmentation outputs for three video lectures from different courses. While some phases of detection were correct or accurate, others were not.

Chapter 3

Proposed Methodology

Our methodology to segment these videos includes various libraries, modules and our own ingenuity. Aforementioned libraries play a crucial role in helping us complete our objective. The completion of our objective is enacted in different stages, all of which we have explained in detail, step-by-step below.

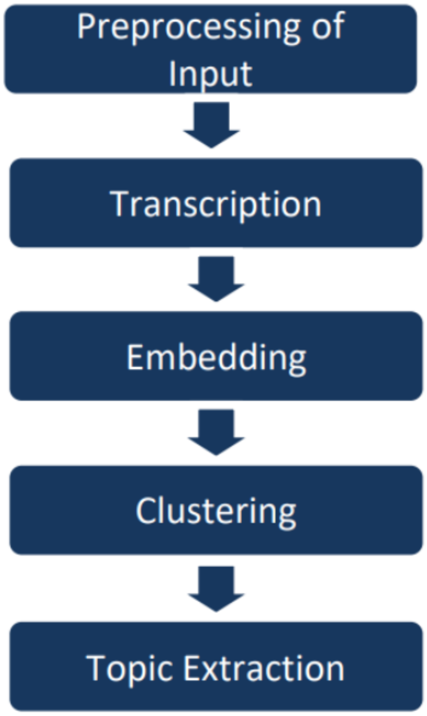


Figure 3.1: Block diagram

3.1 Dataset used

The AVLectures is primarily derived from MIT-OCW. A curated list of courses was formulated by browsing the OCW website. Non-lecture videos were removed and every frame in each lecture was processed and stored using Google Cloud API. Curation of the dataset as a whole was done by separating the dataset into two, CwS (Courses with Segmentation) and CwoS (Courses without Segmentation. The segmentation timestamps are obtained by web scraping. The dataset consists of 86 courses with 2,350 lectures for a total duration of 2,200 hours. In the dataset each course consists of video lectures, corresponding transcripts, OCR outputs for frames, and lecture notes, slides.

3.2 Preprocessing

[6] In preprocessing, the audio cues from the video file are extracted into a .mp3 file. This audio is then split into 10 second chunks using the PyDub library. The segmented audio chunks are again converted into a .wav file and the volume of each chunk is increased to ensure little to no compression and clear audio. Pydub is a library with which we can play, split, merge or edit audio files.

3.3 Transcription

After a deep-dive into the potential modules we could use to transcribe pre-processed audio, we first decided to utilize the SpeechRecognition module. This was not an ideal module to use as the word error was high. It resulted in around 40word error rate. We then discovered OpenAI's Automatic Speech Recognition system which is called Whisper. This was relatively a more ideal pick for transcription as the word error rate when using it was 20more seamless and quicker and overall misinterpretations were all low.. The API itself is based on an Encoder-Decoder model. The input audio is converted into a log-Mel spectrogram which in turn is also passed onto a set of encoders and decoders each of which serves its own function such as caption prediction, language identification, phrase-level timestamps, multilingual speech transcription and even speech translation.

3.4 Embedding

The next step is to embed the transcript sentences of each chunk into vector form. For this the transcribed sentence of each chunk is tokenized into smaller units of words or subwords. The tokens are then passed through a deep neural network model containing multiple transformer layers. This auto-encoder neural network encodes the information or data from the input into fixed-size vectors for each sentence. All this is done using a sentence transformer, where the sentences are passed to get a vectorized embedding. The sentence transformer contains a contrastive loss function which ensures that similar sentences have vectors close in the vector space and vice versa which will help in learning meaningful representations.

3.5 Clustering

The final step in our model was to cluster similar data points together. Once again, we decided to survey our options when it came to this task. Firstly, we decided to eliminate the K Nearest Neighbour as an option simply because it was not providing us with the ideal output, the cluster accuracy was not enough. To improve this, we settled on a clustering methodology called Hierarchical Clustering or Agglomerative Clustering to be specific which is a bottom-up clustering method. It classifies the data points into large clusters first before segmenting them into smaller more accurate clusters.

3.6 Topic Extraction

For getting the proper topic or title for each clusters we used topic extraction methods that organize and understand large collections of text data, by assigning “tags” or categories according to each individual text’s topic or theme. By using this we are trying to obtain the actual topic discussed in a particular segment of the entire audio lecture. For achieving this we used an algorithm known as Latent Dirichlet Allocation (LDA), refer figure:3.2. LDA is utilized for classification of text to a specific topic. Each document is considered as a multinomial distribution of topics. Here, each topic is once again modelled as another multinomial distribution. LDA assumes all texts given to it are all related. Therefore, choosing the right text is important. We have solved this issue using the Agglomerative Clustering Algorithm which clusters similar chunks of text together. It also operates on the second assumption that all documents contain a mixture of topics. The topic words are then generated based on the probability distribution of the words in the document. The result of this is a pool of words attained by taking the probability distribution of the words in the text chunk. These max valued words in terms of probability would be our title or topic of the lecture.

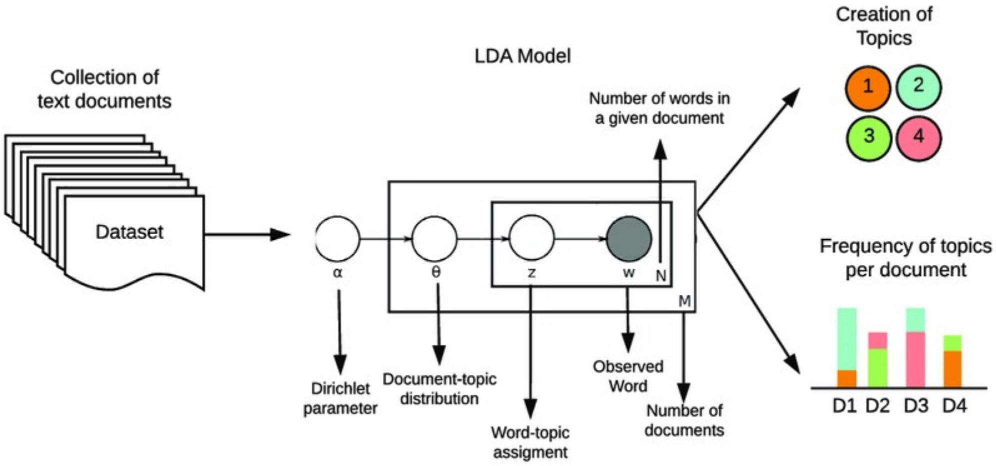


Figure 3.2: LDA Model

Chapter 4

Results and Discussions

4.1 Normalized Mutual Information

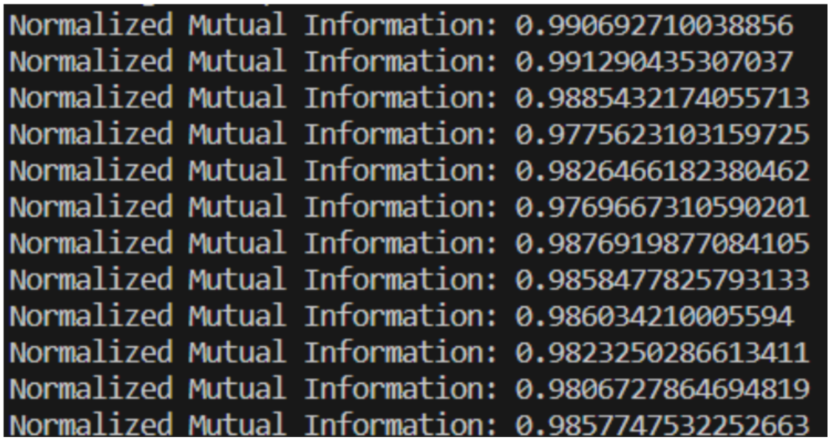
Normalized Mutual Information (NMI) is a metric which mainly measures the similarities between two clusters. As our project involves clustering of audio inputs into clusters which indicate each segment or topic from a particular audio, we think that using NMI we will get a proper idea about the accuracy of segmentation by comparing the clustered segment with the ground truth segments. NMI is a normalization of the mutual information between two clusters, which scales the result between 0 and 1. 0 indicates that there is no mutual information whereas 1 indicates a perfect correlation. NMI can give a proper information of how similar different sets of clusters are, even if the number of clusters differ between the two sets. It is calculated by taking two label arrays as input, which are two different clusters. In our case the clusters are the ground truth segment and the predicted segments.

$$NMI = \sum_{x,y} p(x,y) ln \frac{p(x,y)}{p(x)p(y)}$$

p(x) and p(y) are the marginal probabilities and p(x,y) is the joint probabilities

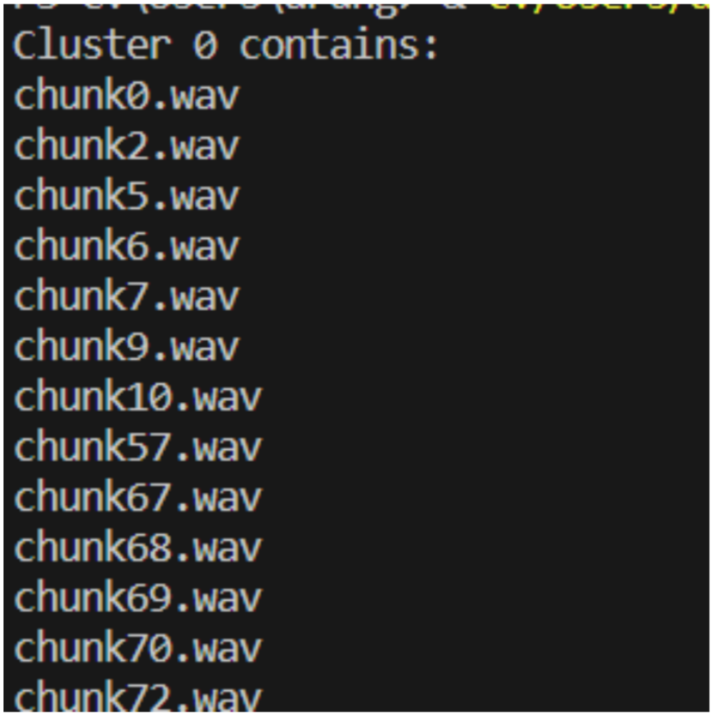
4.2 Result

As a result of the above steps, we will get clustered folders which have audio chunks which belong to the same topic of discussion.



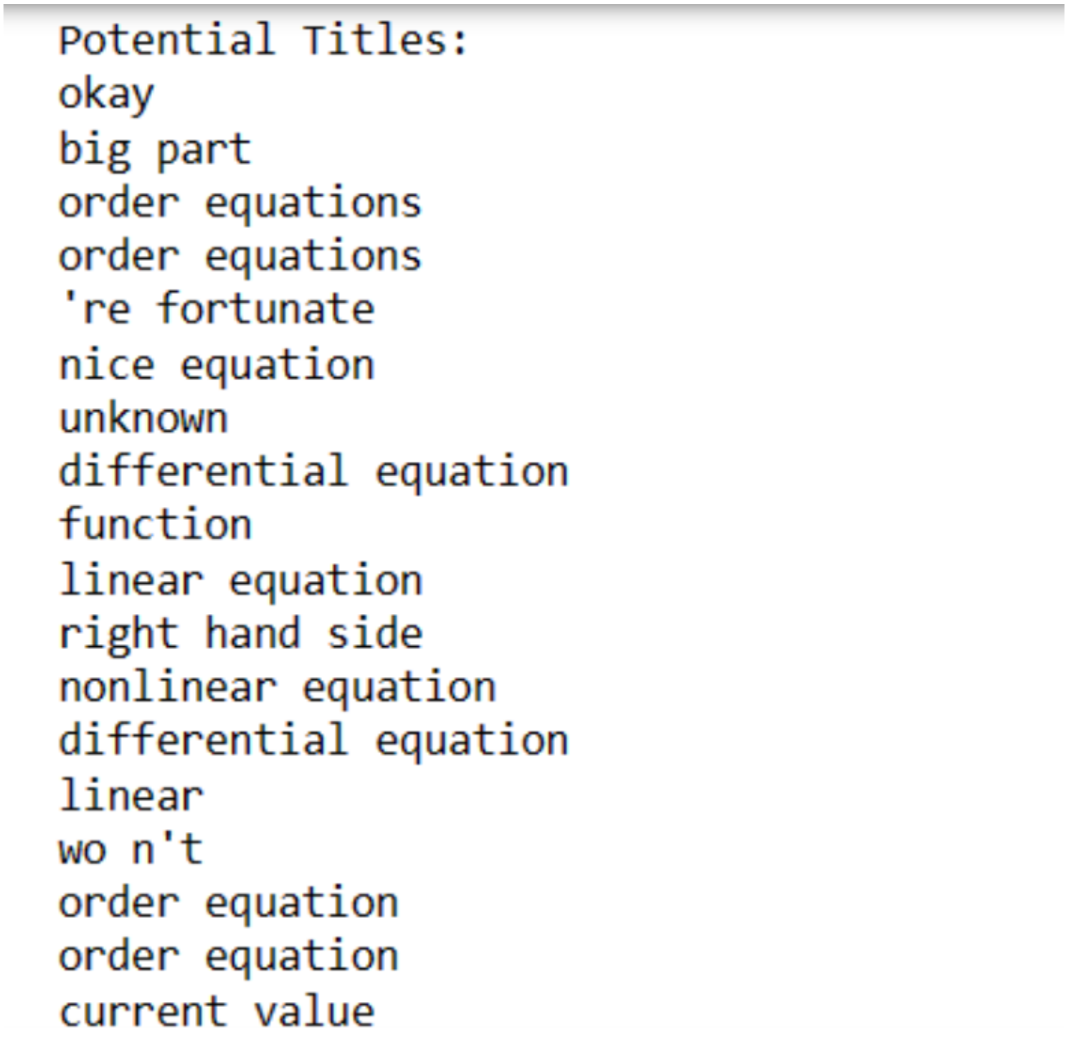
```
Normalized Mutual Information: 0.990692710038856
Normalized Mutual Information: 0.991290435307037
Normalized Mutual Information: 0.9885432174055713
Normalized Mutual Information: 0.9775623103159725
Normalized Mutual Information: 0.9826466182380462
Normalized Mutual Information: 0.9769667310590201
Normalized Mutual Information: 0.9876919877084105
Normalized Mutual Information: 0.9858477825793133
Normalized Mutual Information: 0.986034210005594
Normalized Mutual Information: 0.9823250286613411
Normalized Mutual Information: 0.9806727864694819
Normalized Mutual Information: 0.9857747532252663
```

Figure 4.1: The Normalized Mutual Information values for each clusters in accordance with the ground truth



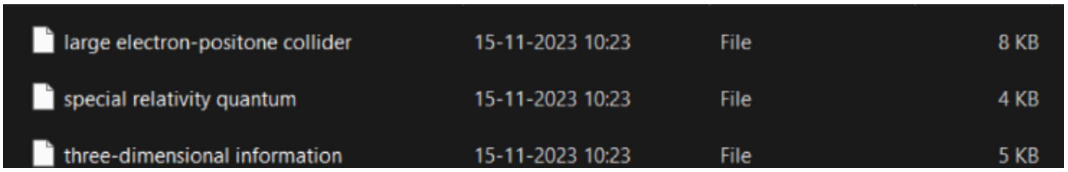
```
Cluster 0 contains:
chunk0.wav
chunk2.wav
chunk5.wav
chunk6.wav
chunk7.wav
chunk9.wav
chunk10.wav
chunk57.wav
chunk67.wav
chunk68.wav
chunk69.wav
chunk70.wav
chunk72.wav
```

Figure 4.2: Example of a cluster after hierarchical clustering which consists of a folder with chunk audio files having similar topics



```
Potential Titles:
okay
big part
order equations
order equations
're fortunate
nice equation
unknown
differential equation
function
linear equation
right hand side
nonlinear equation
differential equation
linear
wo n't
order equation
order equation
current value
```

Figure 4.3: Potential titles obtained for a single cluster using LDA



```
large electron-positone collider   15-11-2023 10:23   File   8 KB
special relativity quantum         15-11-2023 10:23   File   4 KB
three-dimensional information       15-11-2023 10:23   File   5 KB
```

Figure 4.4: Topics saved with title for each segment

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Our objective of segmenting audio is a unique and useful venture. The usefulness of a model that transcribes the audio extracted from a lecture video and groups them together based on their similarity would be incredible. And so we have developed a unique model by utilizing several libraries, modules and an ASR. We have accomplished what we have set out to do, we have applied our model to several lectures present in the AVLectures dataset and we've quantified the accuracy of our model calculating the Normalized Mutual Information. We also plan on improving the model in various ways in our future works.

5.2 Future Work

The future works we will be basing off our current project will also incorporate a visual element not just an audio element. Analysing the videos, frame-by-frame using tools such as Optical Character Recognition will help us in this endeavour. Utilizing the visual component along with the audio component will ensure more dimensionality as the model would be working off two separate criteria.

References

[1] Ankit Gandhi, Arijit Biswas, and Om Deshmukh. Topic transition in educational videos using visually salient words. In *Educational Data Mining*, 2015.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV 2017*, 2017.

[3] Debabrata Mahapatra, Ragunathan Mariappan, and Vaibhav Rajan. Automatic hierarchical table of contents generation for educational videos. In *Companion Proceedings of the The Web Conference 2018*, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[4] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020.

[5] Darshan Singh S, Anchit Gupta, C. V. Jawahar, and Makarand Tapaswi. Unsupervised audio-visual lecture segmentation, 2022.

[6] jiaaro. Pydub. In *https://github.com/jiaaro/pydub*.