2023 4th International Conference for Emerging Technology (INCET)
Belgaum, India. May 26-28, 2023

Video Summarization using Speech Recognition and Text Summarization

Tirath Tyagi, Lakshaya Dhari, Yash Nigam, Renuka Nagpal
Department of Computer Science and Engineering, Amity School of Engineering and Technology
Amity University Uttar Pradesh, India
Email: tirath.tyagi@gmail.com,lakshaya.dhari@gmail.com,nigamya24@gmail.com,rnagpal1@amity.edu

2023 4th International Conference for Emerging Technology (INCET) | 979-8-3503-3575-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/INCET57972.2023.10169901

***Abstract*—Videos on the internet have been increasingly becoming the chief source of knowledge and information in today's digital age. However, with increasing length of videos and diminishing time to spare in everyone's lives, a need has emerged for Video Summarization tools that can provide a good summary about the content of videos without the need to watch videos in their entirety. In this paper, we introduce a two-fold approach to fetch the subject matter of videos through effective summarization. The employed approach comprises of two phases: the first phase involves performing speech-to-text conversion using an Automatic Speech Recognition(ASR) system based on a Convolutional Neural Network(CNN) for generating respective transcripts for input videos while the second phase involves performing Extractive Text Summarization to summarize the text generated by extracting the important information.**

***Index Terms*—Automatic Speech Recognition, Transcripts, Convolutional Neural Network, Extractive Text Summarization**

I. Introduction

Interacting with Audio and video data have become an integral component of our lives, especially since the COVID-19 pandemic. Such data is used in an array of industries including education, business, medicine that rely on vast quantities of audio and video data. However, the work-loaded, fast-paced lives of people often create problems when the data size is considerable. Understandably, students can find it difficult to go though hours of recorded lectures, executives could find it troublesome to record minutes of long and endless meetings due to a number of commitments and constraints and sometimes it is just an annoying task to watch long videos on the internet in order to grasp the subject matter.

To address this problem, it would be beneficial to have tools that can process such videos and provide us with a practical summary without altering the video semantics. In many instances, such a tool would save time and effort and thus, enhance productivity. The summaries generated henceforth, will be of great use in the near future supporting a variety of work and are already used by many top-level organizations. This paper aims to propose a tool for Video Summarization which comprises of two elements: Speech Recognition and Video Summarization. A suitable approach for Video Summarization has been illustrated in figure 1.

Being the easiest and most universal form of communication, speech remains to be the chief source of communication for people. Speech Recognition allows a machine to realise phrases spoken by humans which in many instances are used

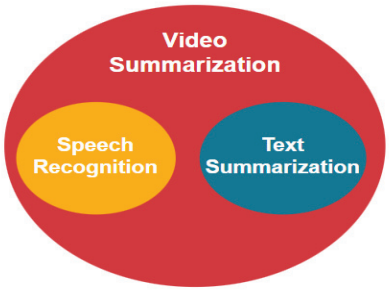

Fig. 1. The Video Summarization process

to generate texts. Speech recognition technology has advanced to the level where it can be used to make machines perform tasks with a wide range of applications in robotics, automation, aviation and computers. Work on speech recognition technology has been going on for a while. Recent advancements in AI and machine learning algorithms have had a huge impact in recognition of speech. The most common products relying on speech recognition technology are virtual assistants which can perform a variety of tasks such as making lists, opening apps and files and so on. Such tools are vital tools for the differently-abled [1]. However, accuracy and computational speed are very crucial when it comes to speech recognition technology. Ensuring high probability of correct words being recognised becomes more difficult in noisy environments [2-3]. Hence, speech recognition has been performed using a CNN architecture and supervised deep learning has been used to train the speech recognition model.

Text summarization takes in a textual input and gives a comprehensive and concise summary as output by extracting vital information from the text. Natural Language Processing (NLP)-based algorithms can be used to summarize transcripts and provide with a summary while preserving the semantic meaning of the text [9-11]. Recently, various applications have emerged for text summarization such as social media posts summarization, written literature summarization, research literature summarization, e-mail summarization, legal and biomedical documents summarization and sentiment analysis, a concept which has grabbed a lot of attention by researchers lately [12]. Based on the algorithm applied, text summarization can be classified into abstractive text summarization and extractive text summarization. Paraphrasing and

extrinsic vocabulary is used to generate a summary by using abstractive text summarization modelling human behavior. On the other hand, extractive text summarization uses intrinsic vocabulary and captures crucial, precise sentences from documents. The importance of these sentences is measured in terms of their frequency and important score [10]. In the proposed tool, an extractive text summarization approach has been employed to avoid paraphrasing, reduce complexity and create language-independent summaries.

As shown in figure 1, a Video Summarization tool has been proposed by using speech recognition technology to generate video transcripts and summarizing the generated text through extractive text summarization.

The following structure has been followed in this paper: Some prior work on video summarization is depicted in section 2. Section 3 discusses the proposed video summarization tool and the associated methodology. Results and analysis are contained within section 4. Section 5 concludes the study based on experimental results and also discusses future prospects.

II. Related Work

In recent years, researchers have realized the need for Video Summarization and deemed it a highly useful and rather necessary tool with applications in industries such as business, medicine, education, law in addition to people's personal lives. This section accounts prior work done to implement video summarization and also discuses related work in speech recognition and text summarization technology.

A detailed study of speech recognition systems by Neha Jain and Somya Rastogi (2019)[1] gives an account of evolution of speech recognition systems and subsequent advancements involving the use of neural networks. CNNs were used on small training sets by V Poliyev and O N Korsun (2020) [2] to perform speech recognition covering four scenarios: speaker-dependent recognition without noise, speaker-independent recognition without noise, speaker-dependent recognition with noise, speaker-independent recognition with noise. With the selection of optimal CNN architecture, good recognition results were achieved with low error-rates in noise-based scenarios. Alsobhani et al. (2021) [3] have been able to use CNNs to perform the feature extraction stage of speech recognition with decreased complexity and greater accuracy. A Deep learning model allows for performing speech recognition in noisy environments with greater accuracy. TensorFlow has been used by Tausif et al. (2018) [4] to train a neural network for developing a model for conversion of speech in Bengali language to text. The trained model gave 95% accuracy for the training set. In spite of limited data availability, test data provided 50% accuracy.

Various approaches and techniques for automatic text summarization have been studied and reviewed by Widyassari et al. (2022) [5]. The most popular approaches can be categorized into extractive and abstractive summarization. Alexandra Savelieva, Bryan Au-Yeung and Vasanth Ramani (2020) [6] propose a model for abstractive summarization using BERT to generate summaries competitive to the human-generated YouTube descriptions. While the BERT model and K-means clustering have been used by Srikanth et al. (2020) [7] to perform extractive summarization and a dynamic method of producing summaries of suitable sizes depending on the size of the cluster has been proposed by the authors.

Recent work shows that NLP-based algorithms also have the potential of performing video summarization. This potential has been explored by Aswin et al. (2021) [9] who have been able to use NLP-based algorithms to generate summaries of videos using their subtitles. In scenarios where subtitles of videos are unavailable, speech-recognition API WIT.AI [8], used by Facebook has been employed to generate subtitles for the videos through speech-recognition and then NLP algorithms such as Lex Rank, Text Rank and Latent Semantic Analysis(LSA) have been used and the results have been compared. Sarah S. Alrumiah and Amal A. Al-Shargabi (2021) [10] develop a summarization model for educational videos. Instead of using the commonly used extractive summarization techniques such as TF-IDF and Latent Semantic Analysis for subtitles summarization, a Latent Dirichlet Allocation(LDA) based approach has been used. In addition, a length enhancement method has been proposed because of lengthy summaries produced by LDA. A hybrid Video Summarization approach has been proposed by Vinnarasu A. and Deepa V. Jose (2019) [11] for performing speech-to-text conversion using Google API and summarization using python NLTK libraries based on sentence ranks by finding work frequencies. Mridha et al. (2021) [12] study work done in the field of Automatic Text Summarization (ATS) and how the existing summarization models generate average summaries that are not always ideal. Combining different approaches, both supervised and unsupervised, was prioritized to generate text summaries that are of higher quality, meet human standards, and are robust. Various phases involved within text summarization like influential text summarization architectures, feature extraction architectures and performance measuring matrices, which would encourage future work in the field of ATS to overcome new challenges in the domain, have been discussed. Shah et al. (2022) [13] talk about the importance of skipping the non-essential parts of a video and the time wasted by jumping on to the most relevant parts of a video. The discussion also includes a coherent technique for video summarization and timestamp creation by using a CNN encoder-RNN decoder neural network. Human evaluation of the generated results showed promising findings. It has been established by Rochan et al. [2018] [14] that although in computer vision, video summarization and semantic segmentation are viewed as separate problems, both these problems share a connection. As a result, semantic segmentation networks has used commonly used for video summarization. Instead of using recurrent neural networks such as long short-term memory (LSTM), using fully convolutional sequence networks (FSCN) have been proposed for video summarization. Obtained results are competitive in performance to the commonly used supervised and unsupervised methods that chiefly rely on LSTMs. A novel technique for automatic production of summaries from multiple documents

has been provided by Zhong et al. (2017) [15]. A four step model based on traditional summary generation algorithms is used to covert multiple documents into a single document which involve using LDA and extended LexRank algorithm and TensorFlow is used on the produced single document to obtain the final summary. A combination of traditional video summarization algorithms are coupled with a deep learning algorithm for performing video summarization from multiple documents. Deep learning has been employed by Ayache et al. (2021) [16] who propose an approach for speech command recognition and use it to make a robot perform particular tasks in a hospital. A convolutional neural network is used for effective feature extraction in place of using conventional approaches.

The methodology for developing the proposed video summarization tool has been discussed in the following section.

III. Proposed Methodology

The proposed approach employs automatic speech recognition(ASR) coupled with text summarization to perform video summarization. A deep learning CNN speech recognition model is used to generate transcripts for videos in scenarios of their absence. The model is a sequence-2-sequence (seq2seq) model based on Connectionist Temporal Classification (CTC) which can be readily used for speech recognition. CTC is the ideal approach because of unaligned datasets common in case of speech-to-text conversion. Word Error Rate (WER) has been used to evaluate model accuracy. The output transcripts in textual form serves as an input to the NLP-based algorithms that provide a concise and informative summary using extractive summarization.

A. Speech Recognition

The first phase of the proposed tool revolves around speech recognition. Speech recognition is performed so as to generate a transcript of what has been said in a video before it can be processed through a text summarization process. To fulfill this purpose, a CTC-based ASR model has been developed. CTC is an algorithm which is used for training deep learning models for seq2seq problems such as speech recognition, handwriting recognition and so on. To develop the model, CTC has been combined with a 2-D Convolutional Neural Network(CNN). For validation and testing, LJSpeech dataset, derived from the LibriVox project has been used. This dataset contains about 14000 recordings. In addition, a custom dataset of 1500 recordings has also been created. WER is used to measure the accuracy of the model. WER is calculated by adding up substitutions, insertions and deletions that occur in a sequence of predicted words. The resultant value is divided by the total number of words that were actually spoken as per the generated transcript. A library called jiwer has been used to calculate the word error rate. Consequently, approximate transcripts of videos are generated using the model. These transcripts act as the input to the text summarization process. The output is being saved inside a temporary text file which will be deleted as soon as the client is served.

1) Challenges: A key problem encountered while performing speech-to-text conversion is the spoken accent. Spoken accents could vary from person-to-person and the dataset used in this study to train the speech recognition model comprises of data having spoken accents only for five different people. Consequently, the model was able to perform speech-to-text conversion accurately only for those five people. Predictions beyond this scope were fairly inaccurate. However, a video summarization tool will be used by a multitude of people and must be able to make accurate predictions for all of them. To bridge this gap, a python library called speech_recognition [17] has been employed which uses pre-trained speech recognition models by Google, Amazon and Facebook to generate transcripts. Finally, a speech recognition system will be established which will be deciding which model to use according to the type of input data received and predictions will be made based on the selected model.

B. Text Summarization

The output from the speech recognition process serves as the input for the second phase, i.e, text summarization. It involves the implementation of a proper text summarization system that can take the previously generated transcript as an input and give out a summarized output either in a form of a sentence or a small paragraph depending upon the length of the input text. NLP-based algorithm called Textrank has been used to perform extractive text summarization in the proposed tool. This is a type of ranking algorithm which is pretty similar to Google's Page Rank algorithm. This is mainly used for extracting keywords and ranking phrases in order to generate a summary. This algorithm measures the relationship between two or more words by using a word probability table or matix as shown in figure 2.

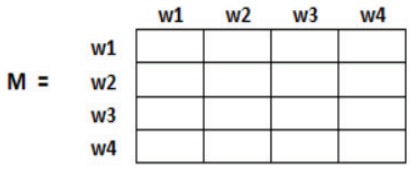

Fig. 2. Word Probability Matrix

For example, if there are four different words w1,w2,w3 and w4; the existing relationship is checked according to the word occurrences by using a table or a matrix. Thereafter, the occurrences are mapped on that particular matrix. This produces a proper matrix as per the text input and that matrix is applied to rank various phrases and find out the probability of more occurrences of words together in a sentence. Essentially, the aim is to find out the importance of every word in the text so as to rank them according to their importance.

A python library generally used for NLP, called nltk is required to divide large text into different phrases. A method under this library called sent_tokenize is used to create a list of

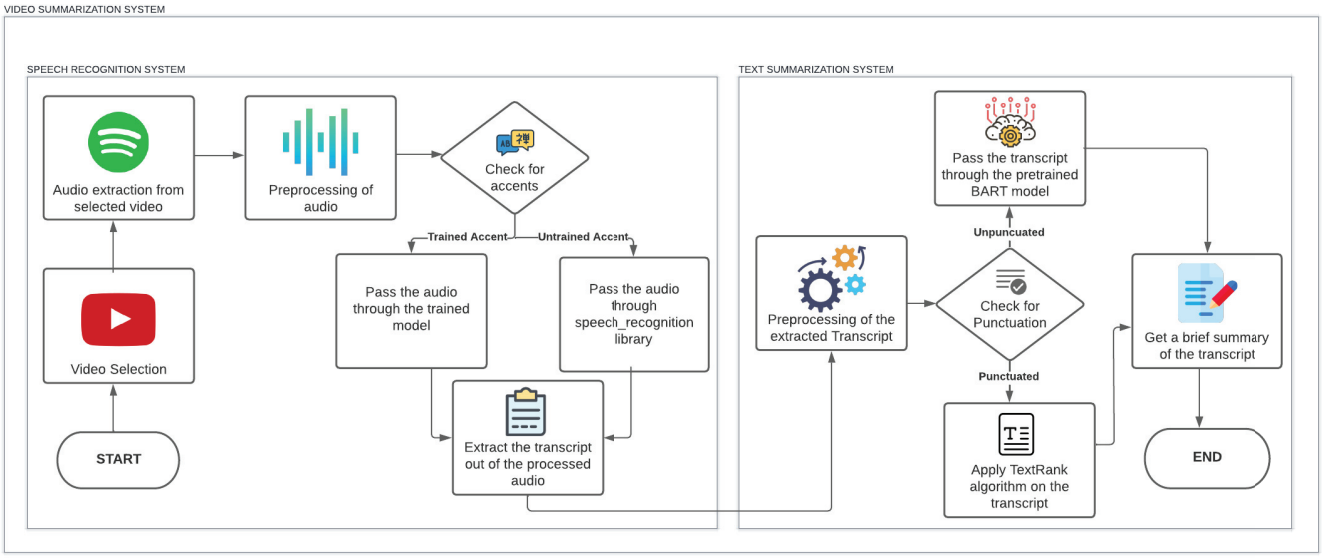

Fig. 3. Proposed Methodology for Video Summarization

sentences in a particular paragraph. Sentences are figured out using the punctuation used in the body of the text. So a period marks the ending of the previous sentence and the beginning of a new sentence. Having figured out the importance of phrases and performed phrase ranking, it can thus be figured out that the text will certainly be about the subject matter of the most ranked phrase and subsequently its relations will be established while traversing down the ranks. Consequently, by using the phrases the target summary can be generated.

The effectiveness of Textrank algorithm can be measured by comparing the word length of the original text and the corresponding summary generated which will verify that word reduction has been performed. Another parameter that can be used to measure algorithm effectiveness is the variation in the semantic meaning of the text.

1) Challenges: While performing text summarization, it is necessary to address the following question: 'What if the transcript is not punctuated?' This is a major challenge as it potentially reduces the proposed tool's functional capabilities. An unpunctuated transcript, if passed through the Textrank algorithm, will give out the text as it is. This is because the sent_tokenize method will consider the whole paragraph as a unit sentence and hence by default, it will be ranked as number one. This challenge can be tackled by introducing a text summarization model which has been pre-trained with a huge dataset. Thus, the Facebook/bart-large-cnn model [18] based on Convolutional Neural Networks(CNN), is employed to generate a summary out of the input transcripts. This eliminates the need for the transcripts to be punctuated. This model has been developed by Facebook and has over 1024 hidden layers and 406 million parameters. The aim is to create a pipeline to this model which can be of aid to retrieve the summarized output for every transcript that is passed through it. Accordingly, a suitable text summarization system will put in place which will be checking for punctuation in transcripts using a separate method and based on that output, the system will select between using the the Facebook/BART-large-CNN model and the Textrank algorithm.

However, using the the Facebook/bart-large-cnn model comes with its own challenges. The problem with pre-trained models from Hugging Face is that they can only accept a limited amount of tokens for predictions. With this model, the tokens were limited to 1024 at a given time. Tokens generally imply number of words. As a solution, long texts can be divided into different parts of fixed word length and which in turn, are then passed onto the model. This will produce a list of summaries for each part being processed which can be ultimately appended together to form one combined summary for the original text.

The detailed process flow for the proposed video summarization is shown in figure 3.

IV. Results and Discussion

Using CNN and deep learning algorithms have proved to be the most suitable approach for solving the speech recognition problem as it allows for effective extraction of features from speech while reducing complexity at the same time. Textrank algorithm is suitable for performing text summarization using extractive summarization. Language-independent summaries have been generated through the proposed tool since extractive summarization relies on in-text vocabulary.

This section includes results and analysis of video summarization after having followed the proposed methodology to develop a video summarization tool for users.

A. Speech Recognition

The speech recognition model has been trained for 50 epochs on a combined dataset comprising of LibriVox data and custom data and for 50 epochs on the custom dataset. GeForce RTX 3090 GPU has been used to train the model with each epoch cycle taking 4-5mn. A sample from our dataset has been illustrated in figure 4.

	file_name	normalized_transcription
0	LJ001-1781	to suggest something that would literally kill...
1	LJ001-0211	The answer to that question was obvious. Peter...
2	LJ001-0332	All he wanted was a candy bar. It didn't seem ...

Fig. 4. Dataset sample

The first column contains the names of the audio files while the second column contains the normalised transcriptions which is nothing but a cleaner version of the transcriptions. All the grammatical components of a language cannot be expressed by speech, there are some components like punctuation and slangs that are although used in everyday life but cannot be expressed in the form of speech. So in order to simplify the slangs and remove all punctuation such that the text is consistent with the speech of a person the normal transcription is converted to the normalized transcription.

The dataset along with the corresponding audio file, signals and labels have been visualized in the form of a spectrogram as illustrated in figure 5.

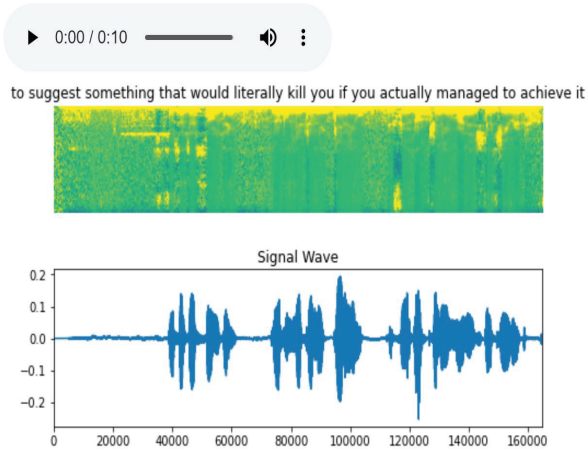

Fig. 5. Spectrogram Visualization

The visual representation showing the variation between the spectrum of frequencies of a signal with time is known as Spectrogram.

The word error rate (WER) for the first training session has been illustrated in figure 6.

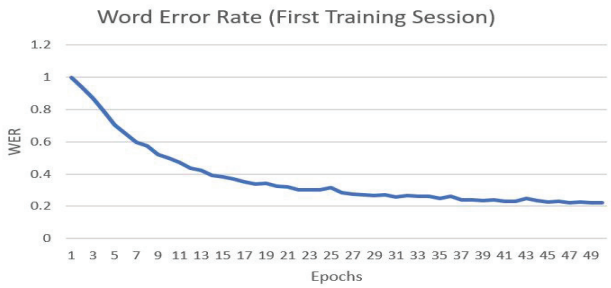

Fig. 6. WER for first training session

The WER seems to be following a parabolic path as the epochs increase. It comes to a constant rate of decrement at around the 45th epoch. The lowest WER during this training cycle turned out to be 0.2210.

The word error rate (WER) for the second training session has been illustrated in figure 7.

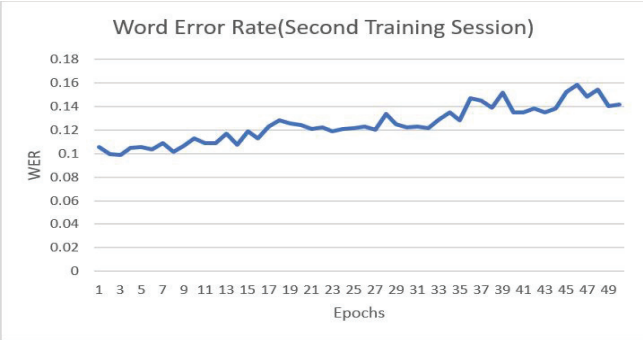

Fig. 7. WER for second training session

The second training session seems to have an almost constant WER throughout the 50 epochs. The variations are generally between 0.10 and 0.15. Taking an average, it can be concluded that the resultant WER is around 0.125. The lowest WER recorded turned out to be 0.10. The illustrated cycle trained the speech recognition model on the custom audio dataset. Testing results showed that the model got used to the authors' voices. The predictions depict that the trained model performed speech-to-text conversion without huge amount of errors after having been trained on the custom dataset.

```
--------------------------------------------------------------------------------
Target    : that is on the order of the mass of the great pyramid of giza in egypt
Prediction: that is on the order of the mass of the creat pyramid of gisa in egypt
--------------------------------------------------------------------------------

--------------------------------------------------------------------------------
Target    : usually explicit feature selection is not the best approach
Prediction: usually explicit feature selection is not the best approach
--------------------------------------------------------------------------------
```

Fig. 8. Predictions made using the Speech Recognition Model

The validation split is derived from the dataset itself. 10% of the dataset has already been split as a validation dataset. Samples from this validation dataset were passed to obtain model predictions as illustrated in figure 8.

The target specifies the actual transcript of the audio file while the prediction is the generated transcript by the trained model.

The problem of spoken accents has been addressed as discussed previously. Applied speech_recognition python library has successfully generated transcripts for unque accents using pre-trained speech recognition models by Google, Amazon and Facebook. An Automatic Speech Recognition (ASR) system has been successfully established which decides upon which model to select according to the type of input data received to make accurate predictions.

B. Text Summarization

Text Summarization has been successfully performed using the extractive summarization approach. The NLP-based TextRank algorithm employed successfully generates target summaries by ranking phrases. However, the TextRank algorithm generates accurate summaries only for punctuated data as predicted previously. Hence, this problem has been addressed and for those use cases wherein the input is unpunctuated, Facebook/bart-large-cnn model has been successfully employed for generating target summaries. A sample input passed through both the text rank algorithm and Facebook/bart-large-cnn model has been illustrated in figure 9.

This Video Summarization tool can be used by students for effectively summarizing course videos. Moreover, it can also be used by working professionals for generating minutes of meetings. It can also be used by people in their day-to-day lives to summarize videos over the internet to gain knowledge and insights. This tool is also highly beneficial for deaf people. We intend to extend our work to implement Sentiment Analysis(SA) in the tool which would provide users with analysis about the embedded emotional tone. Our future work also includes fine-tuning our system. Hence, we plan to introduce a feature in our web application which provides our clients with random phrases and paragraphs which they can read and record (if they wish to contribute). This will increase the size of our current dataset by exponential amounts which can be used to fine tune our model, achieve greater accuracy and introduce various new accents to the model. This will allow the model to make more precise and accurate predictions.

Fig. 9. Input for Text Summarization

The two summaries hence generated have been compared and the output variations based on these summaries have been illustrated in figure 10 and figure 11.

this will increase the size of our current dataset by exponential amounts which can be used to fine tune our model achieve greater accuracy and introduce various new accents to the model.it can also be used by people in their davtoday lives to summarize videos over the internet to gain knowledge and insights

Fig. 10. Output of the Textrank Algorithm

The predictions made by the Textrank algorithm are shown in figure 10. This output is smaller in size as compared to the model predictions but the accuracy is lesser than those of the model.

This Video Summarization tool can be used by students for effectively summarizing course videos. We intend to extend our work to implement Sentiment Analysis(SA) in the tool which would provide users with analysis about the embedded emotional tone. Our future work also includes fine-tuning our system. We plan to introduce a feature in our web application which provides our clients with random phrases and paragraphs which they can read and record. This will increase the size of our current dataset by exponential amounts

Fig. 11. Output generated by Text Summarization Model

Figure 11 shows the output generated by using the facebook/bart-large-cnn model. Even though the length of the summary is greater than the one produced by the TextRank algorithm, the generated summary is more accurate, precise and informative.

Based on the experimental results discussed, the problem of unpunctuated use-cases has been addressed and a pipeline to the the facebook/bart-large-cnn model model has been successfully created to retrieve the summarized output for every transcript that is passed through it. Consequently, a suitable text summarization system has been successfully put in place which checks for punctuation in transcripts using a separate method and based on results obtained, selects between using the the Facebook/BART-large-CNN model or the Textrank algorithm.

Having performed speech recognition and extractive text summarization successfully, the primary objective of summarizing video content is achieved.

V. Conclusion and Future Scope

The proposed approach successfully couples Speech Recognition with Text Summarization to perform Video Summarization. The proposed Automatic Speech Recognition(ASR) system successfully converts speech-to-text for input videos. Depending upon the type of input videos, appropriate transcripts as output are generated which in turn, serve as the input while summarizing text. The proposed Text Summarization system takes in the produced text as input and generates an appropriate and informative summary based on the type(punctuated/unpunctuated) and length of the input text. This Video Summarization tool can be used by students and working professionals. This tool is also highly beneficial for deaf people.

For users to conveniently use the proposed tool, a web-based application needs to be developed. The user-interface would allow users to efficiently summarize Youtube and custom videos. The tool can be extended to implement Sentiment Analysis(SA) which would provide users with an analysis about the embedded emotional tone. Future prospects also include fine-tuning the speech recognition model by using an exponentially larger dataset to introduce multiple accents and allow the model to make more precise and accurate predictions.

REFERENCES

[1] Neha Jain, Somya Rastogi (2019). Speech Recognition Systems - A Comprehensive Study Of Concepts And Mechanism. Acta Informatica Malaysia, 3(1):01-03.

[2] A V Poliyev and O N Korsun 2020. Speech Recognition Using Convolution Neural Networks on Small Training Sets IOP Conf. Ser.: Mater. Sci. Eng. 714 012024.

[3] Ayad Alsobhani et al 2021. Speech Recognition using Convolution Deep Neural Networks. J. Phys.: Conf. Ser. 1973 012166.

[4] M. T. Tausif, S. Chowdhury, M. S. Hawlader, M. Hasanuzzaman and H. Heickal, "Deep Learning Based Bangla Speech-to-Text Conversion," 2018 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII), Yonago, Japan, 2018, pp. 49-54, doi: 10.1109/CSII.2018.00016.

[5] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi, Review of automatic text summarization techniques & methods, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 4,2022, Pages 1029-1046, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.05.006.

[6] Alexandra Savelieva, Bryan Au-Yeung, Vasanth Ramani (2020). Abstractive Summarization of Spoken and Written Instructions with BERT. arXiv:2008.09676

[7] Srikanth, Anirudh & Umasankar, Ashwin & Thanu, Saravanan & Nirmala, Jaya. (2020). Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT. 1-5. 10.1109/ICCCS49678.2020.9277220.

[8] J. Constine, "Speech recognition using wit.ai." [Online]. Available: https://techcrunch.com/2015/01/05/facebook-wit-ai/

[9] V. B. Aswin, M. Javed, P. Parihar, K. Aswanth, C. R. Druval et al., "NLP-driven ensemble-based automatic subtitle generation and semantic video summarization technique," in Advances in Intelligent Systems & Computing, vol. 1133, Singapore: Springer, pp. 3–13, 2021

[10] Alrumiah, Sarah & Al-Shargabi, Amal. (2021). Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement. Computers, Materials and Continua. 70. 6205-6221. 10.32604/cmc.2022.021780.

[11] A, Vinnarasu & Jose, Deepa. (2019). Speech to text conversion and summarization for effective understanding and documentation. International Journal of Electrical and Computer Engineering (IJECE). 9. 3642. 10.11591/ijece.v9i5.pp3642-3648.

[12] Ph. D., M. & Lima, Aklima & Nur, Kamruddin & Das, Sujoy & Hasan, Mahmud & Kabir, Md. (2021). A Survey of Automatic Text Summarization: Progress, Process and Challenges. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3129786.

[13] Shah, Dhiraj & Dedhia, Megh & Desai, Rushil & Namdev, Uditi & Kanani, Pratik. (2022). Video to Text Summarisation and Timestamp Generation to Detect Important Events. 1-7. 10.1109/ASIANCON55314.2022.9909008.

[14] Rochan, M., Ye, L., Wang, Y. (2018). Video Summarization Using Fully Convolutional Sequence Networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11216. Springer, Cham. https://doi.org/10.1007/978-3-030-01258-8_22

[15] Y. Zhong et al., "An Improved LDA Multi-document Summarization Model Based on TensorFlow," 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), Boston, MA, USA, 2017, pp. 255-259, doi: 10.1109/ICTAI.2017.00048.

[16] M. Ayache, H. Kanaan, K. Kassir and Y. Kassir, "Speech Command Recognition Using Deep Learning," 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), Werdanyeh, Lebanon, 2021, pp. 24-29, doi: 10.1109/ICABME53305.2021.9604862.

[17] "Speech Recognition PyPI for other accents"[Online].Available: https://pypi.org/project/SpeechRecognition/.

[18] Facebook/Bart-Large-CNN hugging face model for text summarization. Available at: https://huggingface.co/facebook/bart-large-cnn.