International Journal for Modern Trends in Science and Technology
Volume 10, Issue 03, pages 206-214.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol10issue03.html
DOI: https://doi.org/10.46501/IJMTST1003034

Holistic Video Summarization with Text and Audio

R.Indhuja, P.Abhilakshmitha, S.Joeshika Sebastian, G.Yuva Sivasakthi

Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, Madurai, Tamil Nadu, India.

To Cite this Article
R.Indhuja, P.Abhilakshmitha, S.Joeshika Sebastian, G.Yuva Sivasakthi, Holistic Video Summarization with Text and Audio, International Journal for Modern Trends in Science and Technology, 2024, 10(03), pages. 206-214.https://doi.org/10.46501/IJMTST1003034

Article Info
Received: 02 February 2024; Accepted: 26 February 2024; Published: 03 March 2024.

ABSTRACT

Video summarization is a vital technique for condensing lengthy videos into concise representations that capture the most important content, which integrates visual, textual, and audio data, offers a comprehensive solution for creating more informative and contextually-aware video summaries. In this era of vast multimedia data, the need for effective video summarization methods has become increasingly essential as the lengthening of videos poses a challenge to efficient content consumption. This approach is valuable in a wide range of applications, including video content retrieval, surveillance, and content recommendation systems, where summarizing videos effectively is of paramount importance. In this paper, we present the technical details of our approach and demonstrate its efficacy through empirical evaluation and testing. By extracting audio from the video and combining speech-to-text conversion (STTC) with extractive summarization methods, our paper delves into the technical intricacies of STTC and summarization algorithms and provides comprehensive summaries that capture auditory elements of the videos and provide effective solutions. Also, the summarized text is obtained as audio summary too.

Keywords: STTC-Speech-To-Text-Conversion, DistlBartCnn, Automatic Speech Recognition

1. INTRODUCTION

In today's digital world, the internet and other multimedia platform serves as a vast reservoir of knowledge and information, with videos emerging as a primary medium for content consumption. However, as the length of videos continue to increase, individuals are faced with the challenge of allocating time to watch them in their entirely. This problem has emphasized the need for efficient video summarization capable of filtering the essence of videos and providing concise insights without the need for prolonged viewing. In response to this demand, we present an approach aimed at extracting the subject matter of videos through effective summarization techniques.

The video summarization strategy implemented in this project follows a meticulous two-phase methodology. In the initial phase, Automatic Speech Recognition (ASR) supported by Convolutional Neural Networks (CNNs) is employed for the intricate task of speech-to-text conversion. This involves the generation of detailed transcripts for the input videos, aligning with the fundamental principles of acoustic modeling. Acoustic modeling establishes statistical relationships between linguistic segments of audio signals and phonemes, providing a foundational understanding of the spoken

content [1]. Concurrently, the second phase integrates Extractive Text Summarization, a sophisticated technique adept at distilling key information from the generated transcripts. Integral to the success of the project are technologies such as TensorFlow, a robust open-source low-level library explicitly designed for constructing and training neural networks [2].

The architectural significance of Convolutional layers within the neural network framework cannot be understated. These layers facilitate the production of feature maps, each intricately connected to subsequent sub-sampling layers, thereby enhancing the network's prowess in discerning pivotal information essential for the video summarization process [4]. Emphasizing the importance of preprocessing, the restoration of sentence segmentation and punctuation in the output of the ASR system is paramount for ensuring the structural soundness of the derived text [5]. Beyond these foundational elements, the project integrates ensemble-based techniques. Leveraging speech recognition and Natural Language Processing (NLP)-based text summarization algorithms, this approach significantly boosts the performance of both subtitle generation and video summarization. The ensemble method incorporates two distinct approaches: the intersection method and the weight-based learning method [6].

Complementing these techniques are valuable insights from references that span various dimensions of the project. These include aspects such as video transcription using NLP [9], deep reinforcement learning [10], and a comprehensive exploration of speech recognition technology [11][12][13]. The amalgamation of these references forms a robust, multifaceted foundation for the video summarization project, which extensively utilizes YouTube transcripts. The comprehensive coverage encompasses crucial concepts in speech recognition, deep learning, and Natural Language Processing.

The [14] two components, seq2seq auto-encoder and contrastive learning, are jointly trained through fine-tuning, which improves the performance of text summarization with regard to ROUGE scores and human evaluation. And other text converter that work with the accuracy by the tkinter's GUI [15]works converts the audio neither the video into the text. In addition to [16]MRC, approach to extract relevant opinions and generate both rating-wise and aspect-wise summaries from reviews. Furthermore, [17]This is the result of the adoption of the dynamic base Pythons pyttsx which considers intentionally in adjacent phases of GTTS and AIML, facilitating the establishment of considerably smooth dialogues between the assistant and the users. And [18], it is easy to convert the audio to text in documented format for referencing purposes as it is difficult to scan for the word in the video compared to the transcript.

Other works construct more comprehensive models based on the idea of clustering [19], a unified multimodal transformer-based model which can effectively align and attend the multimodal input. It is important to note that efforts in, [20] RNN is good at temporal dependency modelling, and has achieved overwhelming performance in many video-based tasks, such as video captioning and classification. And We [21] introduce CLIP-It, a language-guided multimodal transformer for generic and query-fouced video. Simultaneously [22] Text summarization is the process of automatically generating natural language summaries from an input videos.

Furthermore, [23] the multi-modal summarization (MMS) from collections of text, image, audio and video. In this work, we propose an extractive multi-modal summarization method that can automatically generate a textual summary and audio summary. And in audio summarization [24] Weights are assigned to words according to the number of occurrences of each word in the text file. This technique is used for producing summaries from the main audio file. Including [25] The audio visual fusion LSTM can exploit the latent consistency between audio and visual information. The self-attention video encoder can capture the global dependency in the whole video stream.

2. LITERATURE REVIEW

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.

Speech Recognition Systems now-a-days use many interdisciplinary technologies ranging from Pattern Recognition, Signal Processing, Natural Language Processing implementing to unified statistical framework. Such systems find a wide area of applications in areas like signal processing problems and many more. The objective of this paper is to present the concepts about Speech Recognition Systems starting from the evolution to the advancements that have now been adapted to the Speech Recognition Systems to make them more robust and accurate. This paper has the detailed study of the mechanism, the challenges and the tools to overcome those challenges with a concluding note that would ensure that with the advancements of the technologies, this world is surely going to experience revolutionary changes in the near future.

The algorithm of separate words automatic recognition based on convolutional neural networks is developed and presented in this paper. Distinctive feature of this algorithm is the training on sets consisting of only hundreds or thousands of samples. Therefore, important problem is the selection of optimal architecture for neural network, which was firstly proposed and tested. After that, four different cases for recognition were researched: speakerdependent recognition without noise, speaker-independent recognition without noise, speaker-dependent recognition with noise, speaker-independent recognition without noise. Finally, we analyse the experiment results that showed good results for all cases of interest.

The use of a speech recognition model has become extremely important. Speech control has become an important type; Our project worked on designing a word-tracking model by applying speech recognition features with deep convolutional neuro-learning. Six control words are used (start, stop, forward, backward, right, left). Words from people of different ages. Two equal parts, men and women, contribute to our speech dataset which is used to train and test proposed deep neural networks. Collect data in different places in the street, park, laboratory and market. Words ranged in length from 1 to 1.30 seconds for thirty people. Convolutional Neural Network (CNN) is applied as advanced deep neural networks to classify each word from our pooled data set as a multi-class classification task. The proposed deep neural network returned 97.06% as word classification accuracy with a completely unknown speech sample. CNN is used to train and test our data. Our work has been distinguished from many other papers that often use ready-made and fairly consistent data of the isolated word type. While our data are collected in different noisy environments under different conditions and from two types of speech, isolated word and continuous word.

Summarization of speech is a difficult problem due to the spontaneity of the flow, disfluencies, and other issues that are not usually encountered in written texts. Our work presents the first application of the BERTSum model to conversational language. We generate abstractive summaries of narrated instructional videos across a wide variety of topics, from gardening and cooking to software configuration and sports. In order to enrich the vocabulary, we use transfer learning and pretrain the model on a few large crossdomain datasets in both written and spoken English. We also do preprocessing of transcripts to restore sentence segmentation and punctuation in the output of an ASR system. The results are evaluated with ROUGE and Content-F1 scoring for the How2 and WikiHow datasets. We engage human judges to score a set of summaries randomly selected from a dataset curated from HowTo100M and YouTube. Based on blind evaluation, we achieve a level of textual fluency and utility close to that of summaries written by human content creators. The model beats current SOTA when applied to WikiHow articles that vary widely in style and topic, while showing no performance regression on the canonical CNN/DailyMail dataset.

This paper proposes an automatic subtitle generation and semantic video summarization technique. The importance of automatic video summarization is vast in the present era of big data. Video summarization helps in efficient storage and also quick surfing of large collection of videos

without losing the important ones. The summarization of the videos is done with the help of subtitles which is obtained using several text summarization algorithms. The proposed technique generates the subtitle for videos with/without subtitles using speech recognition and then applies NLP based Text summarization algorithms on the subtitles. The performance of subtitle generation and video summarization is boosted through Ensemble method with two approaches such as Intersection method and Weight based learning method Experimental results reported show the satisfactory performance of the proposed method.

Speech Recognition Software is a computer program that is trained to take the input of human speech, interpret it, and transcribe it into text. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems. The objective of this paper is to propose an advanced and accurate end-user software system that is able to recognize specific commands to control a robot to perform specified tasks in a hospital. This model will be based on Deep Learning since it is effective in models having huge data as for the two versions of Google TensorFlow and AIY datasets used in our model. Convolutional neural network will be used since it is able to extract features from the dataset instead of traditional methods of feature extraction, thus saving training time and reducing the complexity of the system. With addition to that, NVIDIA CUDA will be also used to train the model with GPU to decrease the training time. During training, some experiments have been done to see the effect of some parameters on the results of the system, and to make sure that the chosen parameters in our model are the best. The results indicate that the training, validation, and testing accuracies of the proposed approach were high, the training duration reached very low values due to the innovation used (CUDA Toolkit) and the commands were successfully recognized by the model. These results outcome the results of the papers that developed similar work which will be presented in the coming sections.

This paper presents a video summarization technique for an Internet video to provide a quick way to overview its content. This is a challenging problem because finding important or informative parts of the original video requires to understand its content. Further-more the content of Internet videos is very diverse, ranging from home videos to documentaries, which makes video summarization much more tough as prior knowledge is almost not available. To tackle this problem, we propose to use deep video features that can encode various levels of content semantics, including objects, actions, and scenes, improving the efficiency of standard video summarization techniques. For this, we design a deep neural network that maps videos as well as descriptions to a common semantic space and jointly trained it with associated pairs of videos and descriptions. To generate a video summary, we extract the deep features from each segment of the original video and apply a clustering-based summarization technique to them. We evaluate our video summaries using the SumMe dataset as well as baseline approaches. The results demonstrated the advantages of incorporating our deep semantic features in a video summarization technique.

This paper proposes a transcript summarization application that works on natural language processing techniques for extracting and summarizing content from audio and video files. The video transcription consists of mainly two parts: first divide the video into several frame-based audio chunks and then the audio chunks are further divided into tokens where each token is then extracted to text. The text obtained is then given to the summarization model. The technique used for summarization is extractive text summarization which extracts summary from top ranked coherent sentences. The efficiency of summarization is evaluated by using videos of different sizes.

Video summarization aims to facilitate large-scale video browsing by producing short, concise summaries that are diverse and representative of original videos. In this paper, we formulate video summarization as a sequential decision making process and develop a deep summarization network (DSN) to summarize videos. DSN predicts for each video frame a probability, which indicates how likely a frame is selected, and then takes actions based on the probability distributions to select

frames, forming video summaries. To train our DSN, we propose an end-to-end, reinforcement learningbased framework, where we design a novel reward function that jointly accounts for diversity and representativeness of generated summaries and does not rely on labels or user interactions at all. During training, the reward function judges how diverse and representative the generated summaries are, while DSN strives for earning higher rewards by learning to produce more diverse and more representative summaries. Since labels are not required, our method can be fully unsupervised. Extensive experiments on two benchmark datasets show that our unsupervised method not only outperforms other state-of-the-art unsupervised methods, but also is comparable to or even superior than most of published supervised approaches.

A real-time system incorporating speech recognition and linguistic processing for recognizing a spoken query by a user and distributed between client and server, is disclosed. The system accepts user's queries in the form of speech at the client where minimal processing extracts a sufficient number of acoustic speech vectors representing the utterance. These vectors are sent via a communications channel to the server where additional acoustic vectors are derived. Using Hidden Markov Models (HMMs), and appropriate grammars and dictionaries conditioned by the selections made by the user, the speech representing the user's query is fully decoded into text (or some other suitable form) at the server. This text corresponding to the user's query is then simultaneously sent to a natural language engine and a database processor where optimized SQL statements are constructed for a full-text search from a database for a recordset of several stored questions that best matches the user's query. Further processing in the natural language engine narrows the search to a single stored question. The answer corresponding to this single stored question is next retrieved from the file path and sent to the client in compressed form. At the client, the answer to the user's query is articulated to the user using a text-to-speech engine in his or her native natural language. The system requires no training and can operate in several natural languages.

In this paper, we present a denoising sequence-to-sequence (seq2seq) autoencoder via contrastive learning for abstractive text summarization. Our model adopts a standard Transformer-based architecture with a multi-layer bi-directional encoder and an auto-regressive decoder. To enhance its denoising ability, we incorporate self-supervised contrastive learning along with various sentence-level document augmentation. These two components, seq2seq autoencoder and contrastive learning, are jointly trained through fine-tuning, w hich improves the performance of text summarization with regard to ROUGE scores and human evaluation. We conduct experiments on two datasets and demonstrate that our model outperforms many existing benchmarks and even achieves comparable performance to the state-of-the-art abstractive systems trained with more complex architecture and extensive computation resources.

3. PROPOSED METHOD

Figure 1 shows the whole flow of methodologies involved in our work. First the audio is extracted from the input video. Then the transcript of the content in the audio is obtained using Speech-to-Text module. The extracted transcript is then converted into user's required language using the multilingual transformation phase. Then with the help of distil-BART-CNN, summarization is done. At last the summary is obtained in audio format using the GTTS python library.

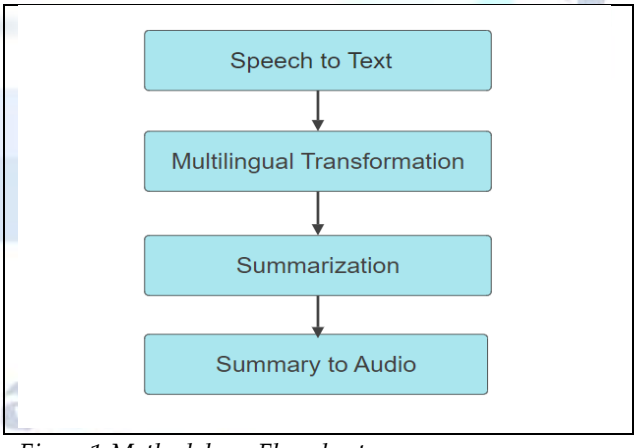

Figure1 Methodology Flowchart

3.1 AUDIO-TRANSCRIPT EXTRACTION

The first phase of our approach deals leveraging a STTC Ssystem to convert speech content within videos into transcripts. The input video is received as a link and the audio is extracted. We use youtube transcript API to

collect subtitle automatically as transcript. In case of unavailable transcripts we use pytube speech recognition model to extract audio and by converting as a wave file to get the content that have spoken in video. The Speech Recognition model accurately transcribes spoken words, capturing the essence of the audio content with high constancy. The wave file extracts the clear audio even the accent that was spoken and collects the transcripts. The extraction of transcript is the foundational view for the summary. The transcripts which are detected, works on the basis of stream chunking form particular seconds and gives the list of files. The list of wave files are provided in the current directory which will be passed ass input like sound file to extract the speech. Once the speech is detected, using transcribe model we can detect speech to text. The Text that is extracted as a transcript automatically generates the content in English language. The most advantageous of its work flow is the "Youtube Transcript API" will be helpful by accepting the other language content as the input and offers us the auto- generated in English language which paves the way to processing of Summary generation.

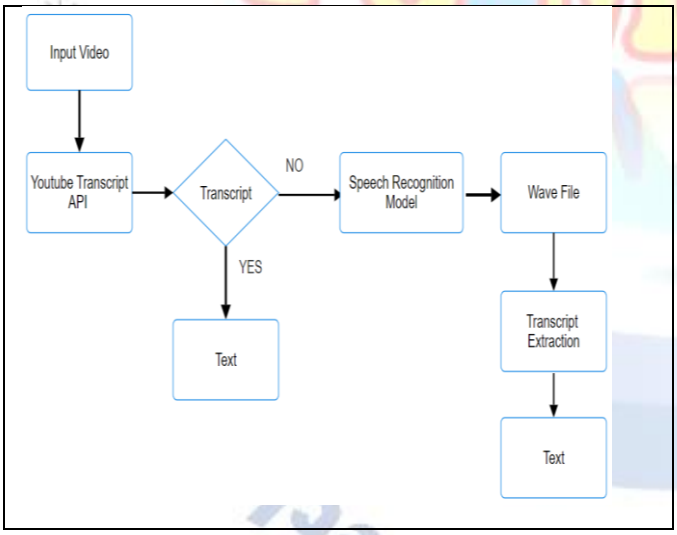

Figure 2 Audio-Transcript Extraction

3.2 MULTILINGUAL MODEL

The main purpose of the multilingual transformation phase is to help the user to understand the video content in their own language. A multi-language model generally refers to a natural language processing (NLP) model that is capable of understanding and processing text in multiple languages. Such models have become increasingly important in the field of NLP due to the global nature of communication and the diversity of languages spoken across different regions. The languages which are available are shown to user to select their particular language with the language code. We used "Googletrans" model, for providing language and language code. The "Googletrans" model is the translator model and the module is available with the multiple languages that are supported by google. The "Googletrans" model is used with the packages like translators and constants which will useful to convert the text for one to another language. The extracted transcripts are passed to the "Googletrans" module to provide the transcripts in the respective language that had been chosen by the user. The translated transcripts are shown to the user, where they can have the knowledge about what does the content of the video is trying to say and informative contents that are been spoken in the provided video

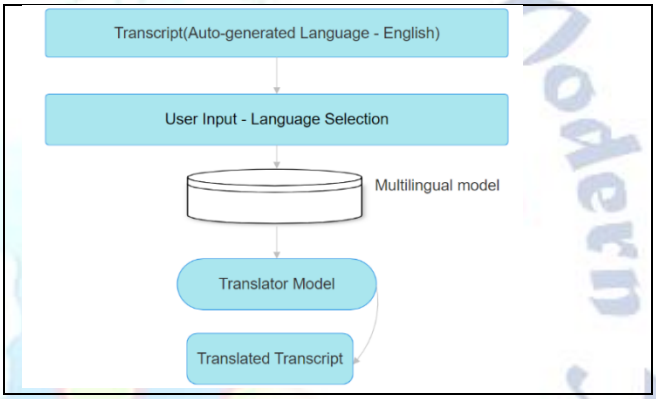

Figure 3 Audio-Transcript Extraction

3.3 SUMMARIZATIONPHASE

The Summary generation is the main outcome and provides the clarity on how the summary generating model is working on. The model which we are using for the summary generation should be applicable to accept the multilingual content in the video and should provide the summary in the multilingual option. The Model which we are using for summary generation is "DistlBART-CNN". Let's see how the model "DistlBart-CNN" works.

INTRODUCTION OF DISTILBART-CNN

DistilBART itself is a distilled version of the BART (Bidirectional and Auto-Regressive Transformers) model, which is a deep learning model based on transformer architecture. BART is based on the transformer architecture, which has proven successful in various natural language processing tasks. The transformer model consists of an encoder-decoder structure, self-attention mechanisms, and multiple layers of neural network components. The Bart and by

extension DistilBART, are both deep learning models used for various natural language processing tasks, such as text summarization. The "distillbart-cnn" has been specifically designed or fine-tuned for tasks related to convolutional neural networks (CNN), it's common to combine the strengths of transformer-based models with CNNs for tasks that involve both text and image modalities.

Working of Distil BART-CNN

Architecture:

BART is a transformer-based model that employs a sequence-to-sequence architecture. It has an encoder-decoder structure where the encoder processes the input sequence, and the decoder generates the output sequence.The model utilizes self-attention mechanisms to capture contextual information from input sequences.

Pre-training:

Like many transformer models, BART is pre-trained on a large corpus of text using auto-regressive language modelling objectives. During pre-training, the model learns to predict the next word in a sequence given the context of previous words. The training phase works and trains with the large dataset which contains phrases of multiple languages.

Fine-Tuning:

After pre-training, the model can be fine-tuned for specific downstream tasks, such as summarization, translation, or question-answering. Experiment with hyper parameters such as learning rate, batch size, and others to find the best configuration for your task.

Summarization:

In the case of summarization, the model is trained to generate concise summaries of input text. The model's decoder is conditioned on the input text, and it generates a summary by attending to important information in the input. As the model got trained with the multilingual phrases, it's easy to generate the summary with multiple languages.

Distillation:

"Distilbart" is a distilled version of BART, meaning it has been trained to mimic the behaviour of a larger BART model but with a smaller number of parameters. This distillation process often involves training the smaller model to reproduce the output probabilities or representations of the larger model on the same tasks.

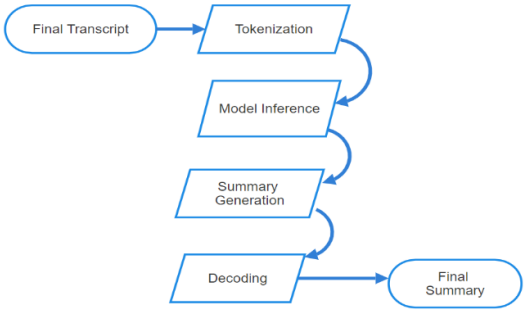

Figure4 Summarization Phase

3.4SUMMARY TO AUDIO

The finally generated summary was shown as a text with concise paragraph where to convert the text summary into audio we used "Google Text to Speech" (GTTS) Model.The gTTS (Google Text-to-Speech) module is a Python library that provides an interface to Google's Text-to-Speech API. It allows you to convert text into spoken words by utilizing Google's TTS service. Google Text-to-Speech often provides multiple voices or accents to choose from. Users may have options to select different voices, languages, and styles. The input text undergoes linguistic analysis to understand the structure, grammar, and pronunciation of words. This involves breaking the text into individual words, identifying punctuation, and determining the appropriate pronunciation for each word. The gTTS module requires an internet connection because it relies on Google's TTS API to generate the speech. The gtts model uses the language code to translate the text to audio that was chosen by the user before. The audio is save with an extension of mp3file. The Audio summary which is done that gives the great impact to people who are wanted to hear instead of reading the texts.

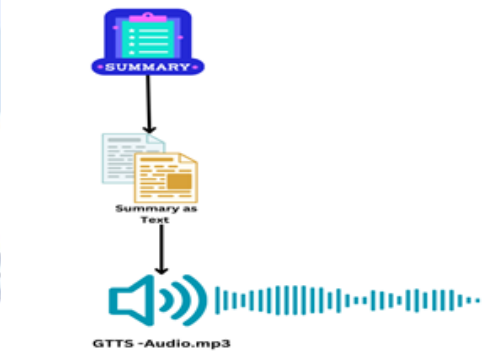

Figure5 Summary to Audio

4. RESULTS AND DISCUSSIONS

Figure 6 shows the language data frame that contains the available language and its language code that could be translated to the user's required language. When the

required language is given as an input, the language's code is returned as an output, using which the particular language is obtained with the help of googletrans API. It is a python library that implements the Google Translate API.

	language	language_code
0	Afrikaans	af
1	Akan	ak
2	Albanian	sq
3	Amharic	am
4	Arabic	ar
..	...	...
120	Western Frisian	fy
121	Xhosa	xh
122	Yiddish	yi
123	Yoruba	yo
124	Zulu	zu

[125 rows x 2 columns]

Figure6 Data Frame

Figure 7 shows the transcript extracted from the audio of the given video input. This is done using the YTapi that is the You Tube API. Using the packages of this python library the audio content in the video gets converted into the transcript form.This Transcript is then given as input the summarization phase.

for Germany it's the end of an era and as Europe's biggest economy there are some huge challenges ahead from its increasingly complex relationship with China to climate change all eyes will be on how Germany's new leaders grapple with these issues but there's one German industry that reveals a lot about the country's prospects it's cars traditionally the car industry here has been very powerful there have been open doors to German Ministries to the Chancery Germany's car industry is a vital part of its economy with links to government that go back decades how this world renowned motor industry navigates the challenges ahead could tell you more than you think about Germany's future [Music] the curves the Precision it's Cutting Edge engineering like this that has helped Place car makers at the heart of the German economy and the industry has proved of Bellwether for the country's future prospects too since the 70s car exports have risen with Germany's wealth and influence manufacturing is incredibly important in Germany...

Fig 7 Extracted Transcript

Figure 8 shows the distil-BART-CNN being trained. This model is especially designed for the summarization. The transcript from the previous phase is given here as input and the model after being trained gives out the summary and figure 9 shows the extracted summary.

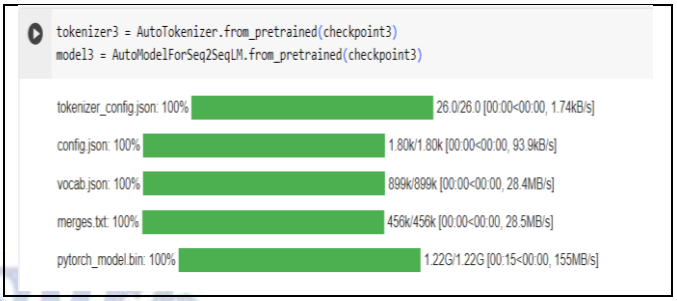

Fig 8 distil-BART-CNN Training

As Europe's biggest economy there are some huge challenges ahead of Germany's new leader . Germany's car industry is a vital part of its economy with links to government that go back decades . Oliver zaipa took over as CEO of BMW's one of the biggest shifts in the company's history in the biggest shift in the history of battery-powered cars in the world's history. Oliver takes on the challenges of the future of the car industry and how this world renowned motor industry navigates challenges ahead …

Figure9 Text Summary

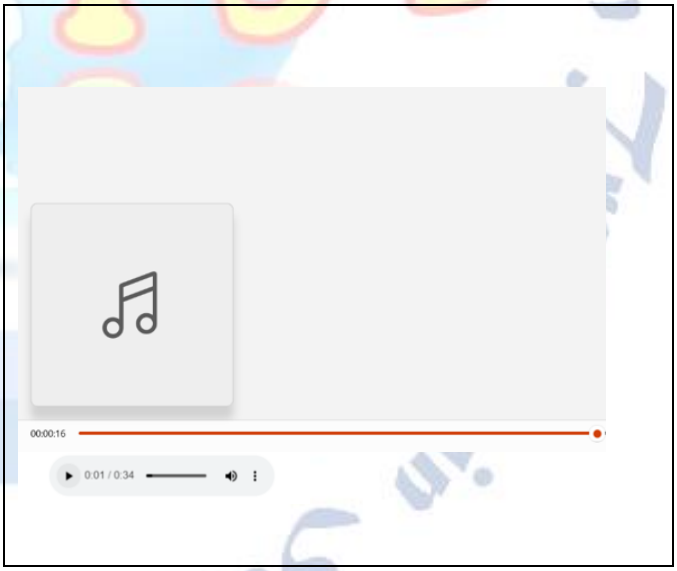

Fig 10 Audio Summary

5. CONCLUSIONS

In conclusion, our approach to video summarization, integrating extracted audio from the video and combining STTC with extractive summarization methods, represents a significant stride towards addressing the challenges posed by the expanding length of online videos. As the digital landscape continues to evolve, the need for tools that filter

meaningful content from lengthy videos has become increasingly apparent. Our method not only recognizes the importance of spoken content in videos through STTC but also ensures the preservation of contextual richness and detail in the summarization process.

The employment of STTC, driven by advanced technologies like distil BART, enables the accurate transcription of spoken words, paving the way for a comprehensive textual representation of video content. This foundational phase is then complemented by Extractive Text Summarization, wherein key sentences and phrases are judiciously extracted to create concise yet informative summaries. By prioritizing the extraction of vital information, our approach maintains fidelity to the original content, offering users a quick and insightful overview without compromising the intricacies of the video's subject matter.

The merits of our approach extend beyond mere efficiency, it reflects a commitment to enhancing the user experience in a time-constrained digital environment. By providing accessible and digestible summaries, we empower users to make informed decisions about content consumption, fostering greater engagement and knowledge acquisition. As online videos continue to proliferate, our approach stands as a valuable contribution to the arsenal of tools aimed at streamlining information retrieval and optimizing the digital viewing experience.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Tirath Tyagi; Lakshaya Dhari; Yash Nigam; Renuka Nagpal, Video Summarization using Speech Recognition and Text Summarization, 2023.

[2] Neha Jain and Somya Rastogi, "Speech Recognition Systems - A Comprehensive Study Of Concepts And Mechanism", Acta Informatica Malaysia, vol. 3, no. 1, pp. 01-03, 2019.

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

[4] Ayad Alsobhani et al., "Speech Recognition using Convolution Deep Neural Networks", J. Phys.: Conf. Ser, vol. 1973, pp. 012166, 2021.

[5] Alexandra Savelieva, Bryan Au-Yeung and Vasanth Ramani, "Abstractive Summarization of Spoken and Written Instructions with BERT", 2020.

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

[7] 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), pp. 24-29, 2021

[8] MayuOtani, Yuta Nakashima,EsaRahtu, JanneHeikkil¨a, and NaokazuYokoya : Video Summarization using Deep Semantic Features In: Proc. Advances in Neural Information Processing Systems (NIPS) 2016

[9] Khushi Porwal, Harshit Srivastava, Ritik Gupta, ShiveshPratap Mall, Nidhi Gupta, Video Transcription and Summarization using NLP,2022

[10] Kaiyang Zhou, Yu Qiao, Tao Xiang:Deep Reinforcement Learning for Unsupervised Video Summarization In: Diversity-Representativeness Reward 2018, Association for the Advancement of Artificial Intelligence.

[11] Rouse, M. Speech Recognition, Available: https://searchcrm.techtarget.com/definition/speech-recognition.

[12] https://www.globalme.net/blog/the-present-future-of-speechrecognition.

[13] Bennett, I.M., Babu, B.R., Morkhandikar, K., Gururaj, P. 2015. Distributed Real-time Speech Recognition, Naunce Communication Inc., Patent No. US 9,076,448.

[14] Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization - Chujie Zheng , Kunpeng Zhang, Harry Jiannan Wang , Ling Fan,and Zhe Wang, 26 Aug 2021

[15] Implementation of Video and Audio to Text Converter - Dr. M. Saraswathi, VVSV Ronit and S Sai Pranav, May 2023

[16] MRCBert: A Machine Reading Comprehension Approach for Unsupervised Summarization- Saurabh Jain, Guokai Tang and Lim Sze Chi, 1 May 2021

[17] An interpretation of AIML with integration of gTTS and Python- Ravivanshikumar Sangpal, Tanvee Gawand, Sahil Vaykar and Neha Madhavi, 05-06 July 2019

[18] Generation Of Transcript In Multiple Languages - Aditya Singh , Dharmesh Saraiya , Rajesh Jethwa , Ms.Mrinal Khasde,2021

[19] Align and Attend: Multimodal Summarization with Dual Contrastive Losses - Bo He, Jun Wang, Jielin Qiu , Trung Bui , Abhinav Shrivastava and Zhaowen Wang ,12 Jun 2023.

[20] Hierarchical Recurrent Neural Network for Video Summarization - Bin Zhao, Xuelong Li and Xiaoqiang Lu,28 Apr 2019

[21] Clip-it! language-guided video summarization - Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell,2021

[22] A Deep Reinforced Model for Abstractive Summarization - Romain Paulus, Caiming Xiong and Richard Socher,2018

[23] Multimodal summarization for asynchronous collection of text, image, audio and video - Li, H Zhu, J, Ma, C, Zhang, J and Zong, C, 2017

[24] Audio Data Summarization System Using Natural Language Processing Pravin Khandare1, Sanket Gaikwad2, Aditya Kukade3, Rohit Panicker4, Swaraj Thamke5, Sep 2019

[25] AudioVisual Video Summarization- Bin Zhao, Maoguo Gong, and Xuelong Li, 17 May 2021