Survey on techniques for Video Summarization and Audio Generation

TANVI KHARE[1], ADITI MOJIDRA[2], PRERNA MAINDARGE[3], SNEHA SALUNKE[4], KALYANI WAGHMARE[5]

[1, 2, 3,4] *Computer Engineering Department, Pune Institute of Computer Engineering, Pune, India.*

[5]*Assistant Professor, Pune Institute of Computer Engineering, Pune, India.*

Abstract- ***The demand for video summarization is rising, driven by the need for quick and efficient access to essential information within voluminous video content. Simultaneously, there's a growing necessity for multilingual audio generation to cater to diverse global audiences. Effective summarization tools are essential, enabling users to extract key insights swiftly. Additionally, integrating multilingual audio capabilities ensures that these summarized contents are accessible to speakers of various languages, enhancing inclusivity. In this paper, we present a comprehensive review of the existing literature on machine learning models and evaluation methods that caters to the requirements of the application. Through an extensive survey of scholarly works, we synthesize and evaluate the key findings, methodologies, and theoretical frameworks in the field. Our review not only offers a comprehensive understanding of the current state of research but also identifies gaps and emerging trends, providing valuable insights for future utilization.***

Indexed Terms- ***Natural Language Processing (NLP), Video summarization, Multilingual text-to-speech synthesis, Keyframe extraction, Recommendation system, Information extraction***

I. INTRODUCTION

In the age of digital media, video content serves as a cornerstone of information dissemination and knowledge sharing. From online tutorials and educational lectures to news broadcasts and entertainment shows, videos offer a wealth of information. However, the sheer volume of video content available online can be overwhelming, making it challenging for users to quickly grasp the essence of a video without investing significant time. In today's fast-paced world, where time is of the essence, our tool revolutionizes the way we interact with video content. Leveraging advanced Natural Language Processing (NLP) algorithms, machine learning prowess, and cutting-edge multilingual Text-to-Speech technology, our Multilingual Video Transcript Summarizer ensures that users can extract key insights, crucial information in their preferred language and the most significant moments from any video, making information more accessible and inclusive than ever before

II. LITERATURE SURVEY

A. Text Summarization

- Different strategies to develop a recommendation system are discussed in [1]. The authors have implemented a movie recommendation system using algorithms such as Alternating Least Squares (ALS),Singular Value Decomposition(SVD),K-Nearest Neighbors (KNN), Co-clustering and Cosine similarity. Cosine Similarity was considered the recommended model based on the accuracy comparison between combination of these models using error measurements.
- In [2] the authors have studied Neural Seq2Seq models for extractive text summarization. However, repetition and inaccurate factual details seem to be obstacles for effective summarization therefore authors have adopted 2 approaches. Hybrid pointer generator network along with coverage helped eliminate the mentioned shortcomings.
- In [3] several extractive summarization strategies like sentence score and summary sentences selection have been discussed. Topic

representation approaches like topic words, Frequency-driven Approaches, Latent Semantic Analysis, Bayesian topic models are briefed. Indicator methods like graph-based approach and machine learning are studied. ROUGE (Recall Oriented Understudy for Gisting Evaluation) model is used to automatically determine the quality Evaluation model by comparing it to human summaries.

- As discussed in [16],Word Graph Methodology, words are separated into two parts. First part is Sentence reduction and then combination of those sentences. Use of Graphs is made where each node represents word information and their relation.It Provides Syntactically correct sentences.It doesn't consider meaning of words.

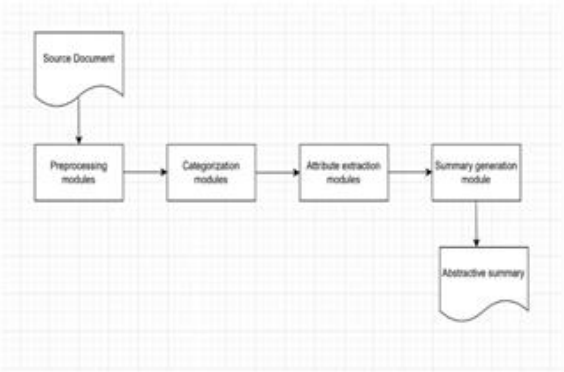

Fig. 1. Abstractive Text Summarization flow

- In [4] three different approaches namely the statistical approach such as Term Frequency Inverse Document Frequency(TF-IDF), the topic modeling approach such as Latent Semantic Analysis (LSA), and graph-based approaches such as TextRank were applied to generate a concise summary for the benchmark the British Broadcasting Corporation (BBC) news articles summarization dataset. The domain specific implementations of each approach in the five domains of the dataset and domain-agnostic prospects were explored in the paper while drawing various insights. The generated summaries were evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework, leveraging precision, recall, and f-measure metrics.
- As discussed in [18], Morkov Clustering Algorithm employs hybrid technique. It forms cluster of sentences which are best-suited and from them it constructs the new sentences. Grouping of sentences is done using Semantic and Statistical variables in this Algorithm to produce closely related sentences. Accuracy depends on quality of reduction techniques. • In [22], Fuzzy Logic has been briefed. It assigns a cost to the sentences in a document and it chooses the phrases based on their relevance which can be obtained by Calculating length of sentence, placement of Sentence, noun and their similarity It can solve unequal weighting of attributes to evaluate their relevance.
- In Pegasus Model (as discussed in [20]), Some irrelevant lines are removed from the input document and compiled as separate output. Here phrases are getting selected on the basis of relevancy and not on randomness. The text generated may require a quick check for errors to increase accuracy of the Summarized text.

B. Multilingual Text-To-Speech

- In [5] the authors have discussed the specifics of the text-to-speech (TTS) system that they have developed.3 sub-types of concatenative synthesis are discussed. They are - Domain-specific Synthesis,Unit Selection Synthesis and Diphone Synthesis. Additionally, Phoneme based speech synthesis and MATLAB based Software implementation is discussed in [7]. The major modules of the text-to-speech system are explained such as Natural Language Processing (NLP) module,Digital Signal Processing (DSP) module,pronunciation module,Prosody Generation etc.

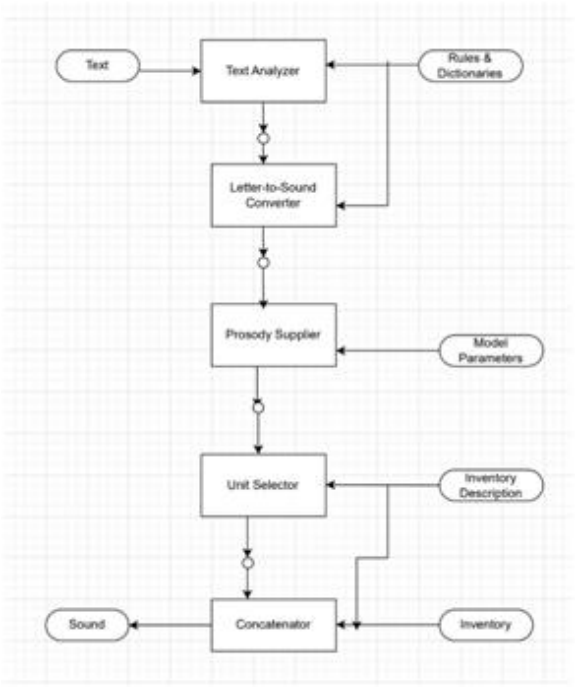

Fig. 2. Text-To-Speech (TTS) Flow

- In [6], a layout is proposed for development of an interactive voice response-based mailing system that enables users to manage their email accounts using audio commands only.Various techniques such as Speech to Text (STT), Text to Speech (TTS), Hidden Markov Model (HMM), Interactive Voice Response (IVR), Mel-frequency cepstral coefficients (MFCC), Dynamic Time Wrapping (DTW), Syllabification,Concatenation, Text Normalization, Text Conversion, Statistical Machine Translation (SMT), Rule Based Machine Translation (RBMT) etc. are summarized. Based on the comparison, the most suitable technique for STT conversion is by deploying a combination of Hidden Markov Model with Deep Neural Network, which can be implemented in Python using Google's Speech Recognition API module. The best method for TTS conversion is by deploying the HMM model that gives the best accuracy and can be implemented in Python using pyttsx3 or gTTS modules. This system of Text-To-Speech and Speech-To-Text can be implemented for to different languages like English, Hindi, Punjabi etc., depending upon the user's requirement and has the capability to recognize these languages and change it to the desired text or speech format.
- Paper [8] presents a model of text analysis for text-to-speech (TTS) synthesis based on weighted finite state transducers, which serves as the text-analysis module of the multilingual Bell Labs TTS system. The paper discusses a text analysis system applied to various languages, including German, Spanish, Russian, Mandarin, and Romanian, with ongoing work on French, Japanese, and Italian. The system utilizes finite state transducers, emphasizing their ability to dynamically construct states during usage, allowing efficient processing without the need for extensive precompilation. The approach challenges traditional methods by integrating operations like word segmentation and numeral expansion within the linguistic analysis phase, contrary to the common preprocessing view. The system's foundation lies in generalized state machines (GSMs), enabling the construction of necessary machine components on-the-fly. This property opens avenues for applications such as discourse analysis and morphological reduplication, showcasing the system's flexibility and potential for linguistic problemsolving. Future work aims to explore GSMs further in various linguistic contexts.
- In [9], a framework is introduced for constructing multilingual text-to-speech systems, focusing on three levels of challenges. Firstly, it outlines the steps needed to create a synthetic voice in a new language from scratch. Secondly, it explores creating a new voice without recording additional acoustic data, considering the limitations of this approach. Lastly, it speculates on the steps required to achieve high-quality synthesis in new languages by recording minimal amounts of data in the target language. The excerpt from the research paper discusses the challenges in building synthetic voices for new languages. While methods have been defined, creating highquality voices still requires expertise and care. Many languages lack synthetic voice support, particularly in regions with low literacy rates beyond major languages. To address this,

researchers are focusing on finding appropriate phoneme sets, understanding speaker-specific pronunciation rules, and improving cross-lingual voice conversion techniques. Ongoing improvements involve adapting from similar languages and refining tools and evaluation methods for easier voice development.

- In [10] the authors present a systematic study addressing the challenge of achieving human-level quality in Text-to-Speech (TTS) systems. They define human-level quality formally and establish guidelines for its evaluation. The study introduces NaturalSpeech, an innovative TTS system, designed to attain human-level quality. The authors identify quality gaps in existing TTS systems and propose several enhancements to bridge these gaps. These enhancements include phoneme pre-training, differentiable durator, bidirectional prior/posterior modules, and a memory mechanism in Variational Autoencoder (VAE). The authors evaluate their Natural Speech system using the LJ Speech dataset, demonstrating that it achieves human-level quality according to CMOS evaluations. Although the system does not surpass or replace human abilities, its quality is statistically indistinguishable from human recordings on the LJSpeech dataset. The authors emphasize that the technologies developed in Natural Speech can be extended to other languages, speakers, and styles to enhance synthesis quality. They express intentions to further pursue human-level quality in challenging scenarios, such as expressive voices, longform audiobook voices, and singing voices, incorporating dynamic, diverse, and contextual prosody in their future work.
- In [12], the emphasis lies in implementing Text Summarization through the Unsupervised Text Rank Algorithm and subsequently transforming the condensed text into an Audio File using the Google Text To Speech API. substantial text chunks or web page URLs are handled as input, generating a summary, and then converting this summary into an audio file with the help of the GTTS API. The paper demonstrates the process of transforming summarized text into an audio file utilizing the GTTS API.

C. Video Highlighting

- In [13], a scalable deep neural network designed for video summarization within content-based recommendation frameworks is introduced. This model predicts the importance scores of video segments by considering both segment-specific and overall video-level information. The research highlights the effectiveness of data augmentation and multi-task learning, addressing limitations arising from dataset constraints. Additionally, to enhance theunderstanding of the video content, action and scene recognition in untrimmed videos are incorporated using state-of-the-art video classification algorithms. Through experiments involving a mix of high-level visual semantic features, audio characteristics, and optical flow analysis, findings emphasize the pivotal role of visual semantic features in the video summarization process.
- In [14] video summarization framework techniques such as keyframe extraction and video skimming are discussed.Uniform Sampling, Image Histogram and Scale Invariant Feature Transform (SIFT) are the main keyframe extraction methods along with VSUMM. K-Means Clustering and Gaussian Clustering are used for shortening the length of videos. The Results for SumMe dataset show that SIFT, VSUMM, and CNN are good approaches to achieve the goal.
- In [15] Video Summarization is applied for 2 datasets - TVSum and SumMe.According to previous studies a Determinantal Point Process (DPP) module increases diversity in summaries, referring to as DPP-LSTM (long short-term memory). This can be improved by addition of adversarial network. However, the authors develop a deep summarization network (DSN) to summarize videos. DSN has an encoder-decoder architecture, where the encoder is a convolutional neural network (CNN) that performs feature extraction on video frames and the decoder is a bidirectional LSTM network that produces probabilities based on which actions are sampled to select frames. Diversity-representativeness (DR) is a reward function that jointly accounts for diversity and representativeness of generated summaries, and does not rely on labels or

userinteractions at all. The DR reward function is inspired by the general criteria of what properties a high-quality video summary should have.

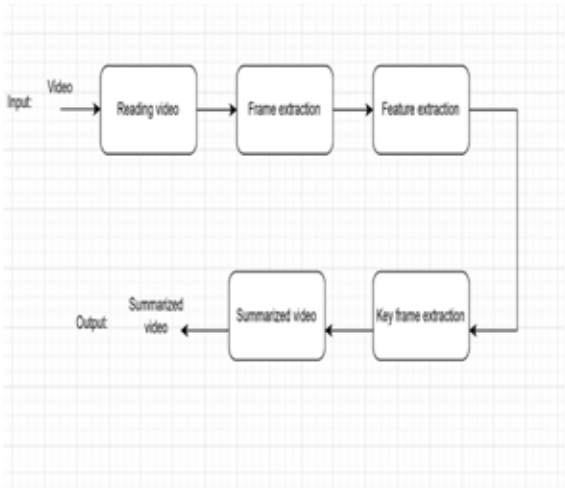

Fig. 3. Video Shortening for Summarization

III. GAPS IDENTIFIED

Identifying gaps is a crucial step as it helps in understanding the specific challenges that the project aims to address. Here are some potential gaps that our Multilingual Video Transcript Summarizer project might address:

A. Information Overload

Users might face difficulty navigating through the videos in search of specific topic. This might require them to skim through the contents of the video quickly. However user might be overwhelmed by the amount of content available.Multilingual Video Transcript Summarizer acts as a filtration system, condensing lengthy videos into concise, informative summaries. By providing users with clear, essential insights, it facilitates efficient content consumption, enabling users to quickly identify and focus on valuable information without being burdened by the sheer volume of videos.

B. Accessibility Challenges

For individuals with disabilities, especially those with visual impairments, accessing video content can be a daunting challenge. Our Text-to-Speech capabilities cater to these accessibility challenges. By transforming textual summaries into clear and engaging audio, our Summarizer ensures that visually impaired users can seamlessly engage with video content. This inclusivity promotes a more accessible digital environment, allowing everyone to benefit from online knowledge.

C. Language Barrier

The language barrier often hampers global access to valuable content. Our Multilingual Video Transcript Summarizer transcends linguistic boundaries. By summarizing videos in multiple languages and extending this facility for audio generation, users can access content in their native language. Whether it's an educational lecture, a tutorial, or a news segment, the Summarizer translates the essence of the content, making knowledge accessible to diverse linguistic communities. This inclusivity fosters a global exchange of ideas and information.

D. Time constraints

In today's fast-paced world, time is a precious commodity. Our Summarizer acknowledges these time constraints by condensing videos into brief, yet comprehensive summaries using abstractive summarization. Users can quickly grasp the main points and critical insights, eliminating the need to invest substantial time in watching lengthy videos.

E. Advertisements

Advertisements often disrupt the flow of video content, causing inconvenience to users seeking relevant information. Our Summarizer employs Segmentation Models to identify and skip advertisement segments, ensuring that the generated summaries focus solely on the substantive content. By filtering out interruptions, users can seamlessly access concise, uninterrupted summaries, enhancing their viewing experience and information retention. F. Personalization Feedback Mechanism based system will help the system to inculcate user's review to improve the performance. The user can decide whether they found the presented information useful or which parts were not up-to-the mark. It acts a personalization method for the user. By understanding user interests and learning from their interactions, the Summarizer not only delivers current summaries but also anticipates user needs and their interest in similar content, ensuring continuous, tailored content recommendations.

CONCLUSION

In our research survey, we delved into the intricate domain of information condensation. We examined

the evolution of text summarization techniques, from extracting key information to generating abstract insights, paving the way for efficient data comprehension. Our exploration extended to groundbreaking advancements in text-to-speech technologies, where natural language interfaces have transformed textual data into accessible audio formats, enhancing inclusivity. Additionally, we investigated the challenges of multilanguage support, emphasizing the development of techniques enabling seamless translation and understanding across diverse linguistic contexts. In the multimedia realm, we explored innovative video highlighting methods, unraveling strategies to distill complex visual content into concise yet informative presentations. Moreover, our research delved into recommendation systems, dissecting collaborative and content-based filtering approaches. These systems, fueled by data-driven algorithms, personalize user experiences, fostering tailored content discovery. Collectively, these studies underscore the transformative power of intelligent information condensation, reshaping how we access, interpret, and share knowledge in our digital age.

REFERENCES

[1] Kalkar, S., Chawan, P., "Recommendation System using Machine Learning Techniques," *IRJET*, vol. 11, no. 9, pp. 2395-0056, Sep. 2022.

[2] See, A., Liu, P. J., Manning, C. D., "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[3] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K., "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[4] Barman, U., Barman, V., Choudhury, N., Rahman, M., Sarma, S., "Unsupervised Extractive News Articles Summarization leveraging Statistical, Topic-Modelling and Graph-based Approaches," *Journal of Scientific and Industrial Research*, vol. 81, no. 9, pp. 952-962, Sep. 2022. DOI: 10.56042/jsir.v81i09.53185.

[5] Isewon, I., Oyelade, J., Oladipupo, O., "Design and Implementation of Text To Speech Conversion for Visually Impaired People," *International Journal of Applied Information Systems*, vol. 7, no. 14, pp. 25-30, Apr. 2014. DOI: 10.5120/ijais14-451143.

[6] Nagdewani, S., Jain, A., "A Review on Methods for Speech-to-Text and Text-to-Speech Conversion," *International Research Journal of Engineering and Technology*, 2020.

[7] Htun, H. M., Zin, T., Tun, H. M., "Text To Speech Conversion Using Different Speech Synthesis," *International Journal Of Scientific And Technology Research*, 2015.

[8] Sproat, R., "Multilingual text analysis for text-to-speech synthesis," *Computing Research Repository - CORR*, vol. 2, pp. 1365-1368 vol.3, 1996. DOI: 10.1109/ICSLP.1996.607867.

[9] Black, A., Lenzo, K., "Multilingual Text-To-Speech Synthesis," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88.*, 2003. DOI: 10.1109/ICASSP.2004.1326656.

[10] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Liu, T. Y., "Naturalspeech: End-to-end text to speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, 2022.

[11] Rand, A., Huda, A., J., A. H., Al-Shakarchy, N., "Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency," *IJEECS*, 2022.

[12] Voleti, S., Raju, C., Rani, T., Swetha, M., "Text Summarization Using Natural Language Processing And Google Text To Speech API," *International Research Journal of Engineering and Technology*, 2020.

[13] Jiang, Y., Cui, K., Peng, B., Xu, C., "Comprehensive video understanding: Video summarization with content-based video recommender design," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[14] Jadon, S., Jasim, M., "Unsupervised video summarization framework using keyframe extraction and video skimming," in *2020 IEEE 5th International Conference on computing communication and automation (ICCCA)*, 2020.

[15] Zhou, K., Qiao, Y., Xiang, T., "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[16] Modi, S., Oza, R., "Review on Abstractive Text Summarization Techniques (ATST) for single and multidocuments," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 1173-1176, 2018.

[17] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., Kabir, M. M., "A Survey of Automatic Text Summarization: Progress, Process, and Challenges," *IEEE Access*, vol. 9, pp. 156043-156070, 2021.

[18] Sahoo, D., Bhoi, A., Balabantaray, R. C., "Hybrid approach to abstractive summarization," *Procedia Computer Science*, vol. 132, pp. 1228-1237, 2018.

[19] Shinde, M., Mhatre, D., Marwal, G., "Techniques and Research in Text Summarization-A Survey," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 260-263, 2021.

[20] Gupta, A., Chugh, D., Katarya, R., "Automated News Summarization Using Transformers," *arXiv preprint arXiv:2108.01064*, 2021.

[21] Gupta, V., Lehal, G. S., "A survey of text summarization extractive techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, 2010.

[22] Kumar, A. K. S. H. I., Sharma, A. D. I. T. I., "Systematic literature review of fuzzy logic based text summarization," *Iranian Journal of Fuzzy Systems*, vol. 16, no. 5, pp. 45-59, 2019.