International Journal Of Engineering Research Excellence And Applied Science (IJEREAS)
Volume-III (Issue 1) – March 2025 ISSN : 3048-5355

Vox2Text: Empowering Seamless Video transcription with NLP Framework

Minakshi Tomer [1†], Tripti Rathee[2*†]

[1]Information Technology, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, 110058, New Delhi, India.
[2*]Information Technology, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, 110058, New Delhi, India.

*Corresponding author(s). E-mail(s): tomer.minakshi@gmail.com;
Contributing authors: rathee.tripti@gmail.com;
[†]These authors contributed equally to this work.

Abstract

Video content has become increasingly prevalent in various domains. However, extracting meaningful information from videos remains a challenging task due to their unstructured nature. This research paper introduces a novel approach to video-to-transcript summarization using Natural Language Processing (NLP) techniques, specifically leveraging the Wave2Vec and Whisper models. Firstly, the audio stream of the video is converted into a time-series representation using the Wave2Vec model. Subsequently, an Automatic Speech Recognition (ASR) system is employed to transcribe the audio embeddings into textual transcripts. Whisper's advanced architecture, based on transformer networks, enables accu- rate and context-aware transcription, even in the presence of background noise or speaker variations. To generate video summaries, the obtained transcripts are further processed using NLP techniques, including text cleaning, tokenization, and sentence segmentation. The resulting text is then subjected to sentence rank- ing algorithms, considering factors such as relevance, coherence, and importance.

The top-ranked sentences are extracted and seamlessly stitched together, forming a concise and representative summary of the video's content.

Keywords: Natural Language Processing, Text Summarization, Extraction, Speech Emotion Recognition, Wave2Vec model, Whisper model, Automatic Speech Recognition (ASR)

1. Introduction

In the realm of human-computer interaction, one of the most recent developments is the recognition of an individual's emotional state [1]

.The fields of artificial intelligence and natural language processing are growing exponentially, as are their applications. This has served many forgeries. Tools to speed up many complex calculations, data extractions, and explorations [2].

In recent years, a number of methods have been put forth that employ sensors either on the users' bodies (such as physiological or inertial sensors) or in their surroundings (such as cameras and microphones). Though in recent years, people have shifted from text to video when they need to express themselves in public[3]

.In this fast-paced world where time is invaluable, people do not have time to watch long videos on the internet. So to refrain from investing so much time, they instead watch them twice as fast to get the impression of what is explained in the video. The amount of videotape data generated every day is

increasing, thus it's critical to optimize the films for quicker recovery and browsing so that drug addicts may choose the most relevant videos to watch based on their preferences [4].

The video content is increasing on the internet rapidly with the increased use of video cameras and the availability or access of video cameras to the masses. Many popular video sharing platforms such as YouTube are receiving 1000+ hours of video per minute and this number is increasing significantly year by year. In addition to this content, there are daily online meetings and events that have become a daily routine since the pandemic. It is clear that people can miss one of these meetings, lectures because of busy work, time conflicts, or other time constraints. To solve this problem, we need a set of tools that can not only convert this audio-video content into text but also qualitatively summarize it without changing its meaning. This helps save a lot of time and the extracted text can be used in different ways. In this paper a technique is proposed that combines audio and text processing to generate concise and coherent summaries of video content. This framework is basically based on a transformer-based pipeline technique for tuning heavily speech-based video files into clean readable text from the audio. Robust speech transcription is now possible like never before with OpenAI's whisper model.

The rest of the paper is organized as follows: The literature survey of various techniques in the area of text summarization and video transcription is mentioned in Section 2. Section 3 introduced the methodology for the proposed framework. Sections 4 presents the result and Finally, the paper is completed with the conclusion and some future scopes in section 5 and 6.

2. Literature Survey

Several studies have explored the recognition of emotions from speech and text using deep neural networks (DNNs) [5]. One such approach proposed by the author is a hierarchical DNN dataset framework capable of recognizing emotions in both single-modality and multi-modality systems. The framework combines 33 acoustic features and their statistical functions to capture information from various levels of speech segments. Additionally, textual features are extracted using ELMo v2 word embeddings, enabling the extraction of contextual and character-based information from text transcriptions. To evaluate the performance of the proposed approach, the author conducted experiments on three widely used datasets: RAVDESS, SAVEE, and IEMOCAP. This research paper [6] introduces a comprehensive automated subtitle generation system composed of three key modules. It serves as a valuable addition to the literature on automated subtitle generation systems, providing insights into the design and implementation of the three key modules. The findings presented in this study contribute to advancing the field and can guide future research in the development of more sophisticated and robust subtitle generation systems.

Event extraction plays a crucial role in natural language processing (NLP) appli- cations [7]. This research paper emphasizes the significance of event extraction and focuses specifically on the annotation of event triggers and arguments that pertain to a specific set of types related to edit actions, including Select, Add, Remove, and Modify. The annotation process aims to contribute to the comprehension and analy- sis of these edit actions within textual data. By conducting this literature survey, we intend to provide a comprehensive overview of the importance of event extraction in NLP research. We explore the fundamental concepts of event extraction, emphasizing the role of event triggers and arguments in capturing essential information. Addi- tionally, we discuss the relevance of annotating edit actions and its implications for understanding text data. Speech Emotion Recognition (SER) is a prominent research area that aims to identify emotions from speech signals [8].

In this literature sur- vey, the author presents novel fine-tuning strategies for wav2vec 2.0 in SER, leading to state-of-the-art (SOTA) performance on the extensively studied IEMOCAP cor- pus. The study highlights the existence of domain shift in SER and emphasizes the importance of addressing this shift to enhance performance. Furthermore, the author introduces an algorithm designed for learning contextualized emotion representation, showcasing its benefits in fine-tuning a wav2vec 2.0 model specifically for SER. These techniques offer potential applications in various other tasks and lay a foundation for investigating the utility of contextualized emotion representation. This literature survey significantly contributes to the understanding of SER by proposing innovative fine-tuning strategies for wav2vec 2.0.

The research paper presents an approach that utilizes a pre-trained speech-to-text Whisper model and incorporates pre-training on synthetic captions [9]. The paper provides an in-depth explanation of the training procedures employed and reports the results of various experiments, including investigations into model size variations, dataset mixtures, and adjustments of hyperparameters. The study findings reveal the significance of leveraging a checkpoint pre-trained on speech-to-text, as it contributes to notable improvements in performance. Moreover, the investigation into pretraining on a combination of synthetic and human-written captions from a different source demonstrates additional enhancements when compared to fine-tuning alone.

Spelling errors are a persistent issue in written communication and can have significant implications for the accuracy and comprehensibility of natural language pro- cessing tasks. Researchers have developed spelling correction systems, with NeuSpell

[2] emerging as a notable open-source toolkit designed specifically for English. This literature survey provides an in-depth analysis of NeuSpell, exploring its features, benchmarking methodology, and novel enhancements that improve contextual under- standing in spelling correction.

The authors in [10] suggested a transcript summary program that extracts and summarizes content from audio and video files using natural language processing techniques.The authors in [11] suggests a method for summa- rizing YouTube videos that preserves the most important parts while using natural language processing (NLP) and machine learning.The recommended approach entails obtaining transcripts from the user-provided video link, followed by employing Pipelin- ing and Hugging Face Transformers to summarize the content. The developed model receives as input from the user video URLs and the necessary summary duration, and outputs a summarized transcript. The authors in [12] have offered a unique method for extracting important linguistic aspects from videos by segmenting lectures and integrating Natural Language Processing tasks.The authors in [13] have proposed an approach to automatically create timestamps and captions for videos.The authors in
[14] have offered a system designed for home users to browse content-based broadcast news videos.

3. Related Work

In this section, we present an overview of the related work in the field of video to tran- script summarization, with a particular focus on our novel approach utilizing Wave2vec and OpenAI's Whisper model. Our work aims to address the challenge of automati- cally generating concise and informative summaries of video content in textual form. By leveraging the power of Wave2vec's audio representation learning and the accurate transcription capabilities of the Whisper model, we strive to achieve accurate and effi- cient transcription of the audio track in videos. Through this combination, we aim to provide a comprehensive solution for video summarization, enabling users to quickly grasp the key information conveyed in videos through concise textual summaries.

3.1 Amazon Transcribe

Amazon transcribe is an automatic speech recognition service that uses Machine Learning ML Models to convert speech into text. With Amazon Transcribe, one can record voice input, create easy to read transcripts, improve accuracy with language adaptation and filter content to ensure customer privacy. Practical use cases include transcription and analysis of customer agent calls and creation of video captions.

3.2 Google Transcibe

Live Transcribe is a real-time caption smartphone application developed by Google that takes speech and turns it into real time caption with simply the usage of the phone mic. It permits two way verbal exchange through a type back keyboard for the users who cannot or do not need to speak.

4. Methodology

Using Natural Language Processing (NLP) techniques, specifically leveraging the Wave2Vec and isper models. The proposed methodology combines audio and text processing to generate concise and coherent summaries of video content. Firstly, the audio stream of the video is converted into a time-series representation using the Wave2Vec model, which captures phonetic and semantic information. This rep- resentation is then transformed into a sequence of acoustic embeddings, providing a foundation for subsequent analysis. Next, the Whisper model, a state-of-the-art Automatic Speech Recognition (ASR) system, is employed to transcribe the audio embeddings into textual transcripts. Whisper's advanced architecture, based on transformer networks, enables accurate and context-aware transcription, even in the presence of background noise or speaker variations. To generate video summaries, the obtained transcripts are further processed using NLP techniques, including text clean- ing, tokenization, and sentence segmentation. The resulting text is then subjected to sentence ranking algorithms, considering factors such as relevance, coherence, and importance. The top-ranked sentences are extracted and seamlessly stitched together, forming a concise and representative summary of the video's content.

4.1 Transcript Generation

Transcription of a video is extracting the texts from a video file or precisely can be said to be a textual form of audio in that particular video. The process of generating transcription starts from an audio file with .wav format which enables us to perform operations on them. Pydub helps us with a function *split$_o$n$_s$ilence*() which splits the audio into small chunks depending on the silences found in the file.

A folder to create the chunks is created so that they can be fetched easily when required to be converted to text. A loop statement is used over the chunk of audio obtained to get in touch with every part of the original audio. The loop would have the heart of the whole process or the most critical thing. The chunks obtained are con- verted and exported into their respective .wav file. An instance of speech recognition, recognition is created to recognize the speech from an audio source. In the loop after every chunk is obtained it is sent to the record function that records the file to the Audio Data instance. The next step being the most important is the final text from that AudioData which is input to the *recognize$_g$oogle*() function which results in the extracted text. This process is repeated and finally, the text obtained is given as the transcript of the video provided.

4.2 Flowchart

This comprehensive flow chart outlines the step-by-step process of converting video files into concise and informative transcripts. The flowchart aims to facilitate seam- less access, analysis, and utilization of video content. The flow begins with the input video file, from which the audio is extracted and

converted to the .wav format. Subse- quently, speech recognition-specific NLP models are employed to transform the audio into text, resulting in a preliminary transcript. To condense the information, text summarization techniques are applied, extracting the most crucial details. The tran- script undergoes further refinement to address typos, inaccuracies, and noisy language, enhancing its accuracy and readability. The output of this flowchart is a clean and condensed transcript, suitable for various applications such as subtitle generation, accessible text indexing, content analysis, and improved accessibility for individuals with hearing impairments. By following this flowchart, researchers and practitioners can efficiently navigate the video-to-transcription summarization process and leverage the potential of this approach.

Steps Followed: Fig 1 represents the Working of Video to Transcript Summariza- tion.
It can be explained in the following:

1. Input Video File: The first step is to supply the video file that is to be converted into a transcript .It could be in formats like .mp4.
2. Extract Audio from Video: The audio is extracted from the video file using a tool like "ffmpeg". This process creates an audio file, such as an .mp3, by separating the audio track from the video.
3. Convert Audio to .wav Format: The extracted audio file must be converted to
4. .wav format. Tools like "ffmpeg" are frequently used to facilitate this conversion. For operations involving audio manipulation, the .wav format is often used.
5. Audio to Text Conversion: In this phase, speech recognition-specific NLP models are used to process the converted .wav audio file. 'Wav2Vec' and 'whisper' are two examples of such models. These models parse the audio input, transform it to text, and foster a preliminary transcript.
6. Text Summarization: The preliminary transcript created in the preceding stage could be long and include superfluous and pointless material. Techniques for text summarizing are used to compress the transcript and extract the most crucial and pertinent information. These techniques can include extractive summarization, which extracts and emphasizes vital phrases.
7. Clean Summarized Transcript: Even after the material was summarized there could still be typos, inaccurate information, or noisy language in the transcript. The transcript is further processed in this phase to make it cleaner. This can entail activities like fixing grammatical issues, spelling mistakes, and improving readability.
8. Output Transcript: The video is cleaned, condensed transcript is the ultimate output. This transcript could be used for a variety of things, such making subtitles, producing accessible text for indexing, assisting content analysis, or rendering the video accessible to those with hearing impairments.

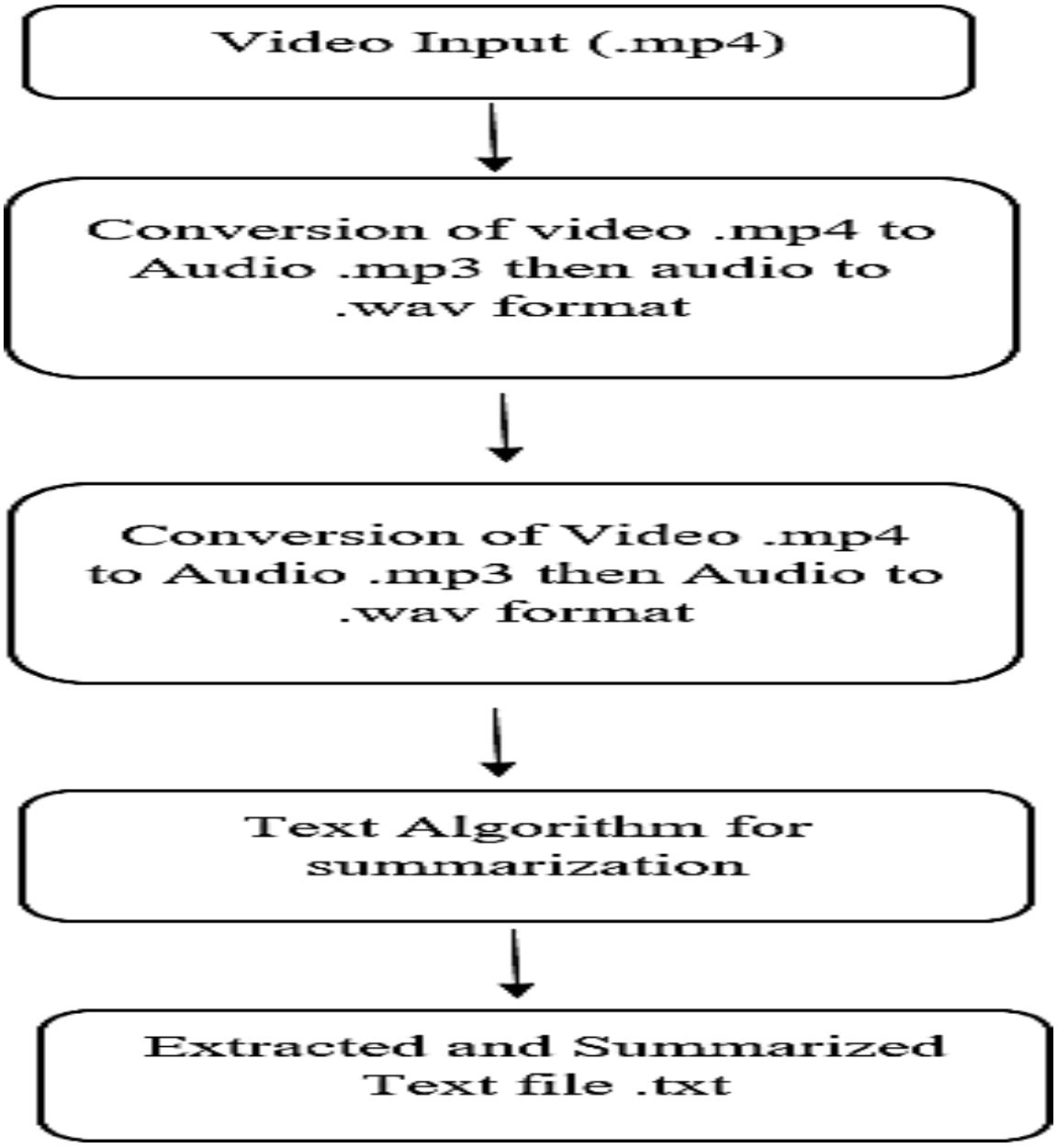

Fig. 1 Working of Video to Transcript Summarization

5. Results

The result section of this research paper presents a comprehensive analysis of video- to-text summarization, focusing on the application of natural language processing (NLP) techniques to generate accurate transcripts from videos. In this study, we com- pare two distinct outputs: the first being the transcript generated by the widely used Whisper model, and the second being the transcript generated by our framework. Our framework incorporates additional filtering mechanisms, utilizing Neuspell for spell checks, ensuring proper punctuation, and employing sequence labelling techniques to enhance the quality and coherence of the output. By evaluating and comparing these

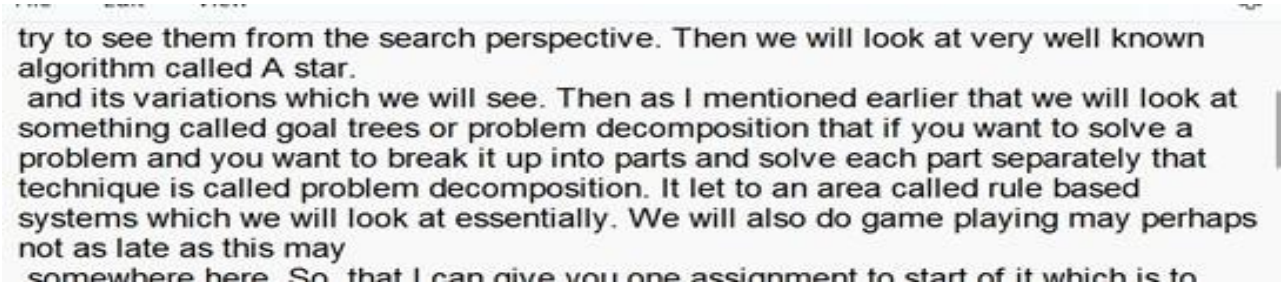
try to see them from the search perspective. Then we will look at very well known algorithm called A star.
and its variations which we will see. Then as I mentioned earlier that we will look at something called goal trees or problem decomposition that if you want to solve a problem and you want to break it up into parts and solve each part separately that technique is called problem decomposition. It let to an area called rule based systems which we will look at essentially. We will also do game playing may perhaps not as late as this may
somewhere here. So, that I can give you one assignment to start of it which is to

Fig. 2 Whisper model output

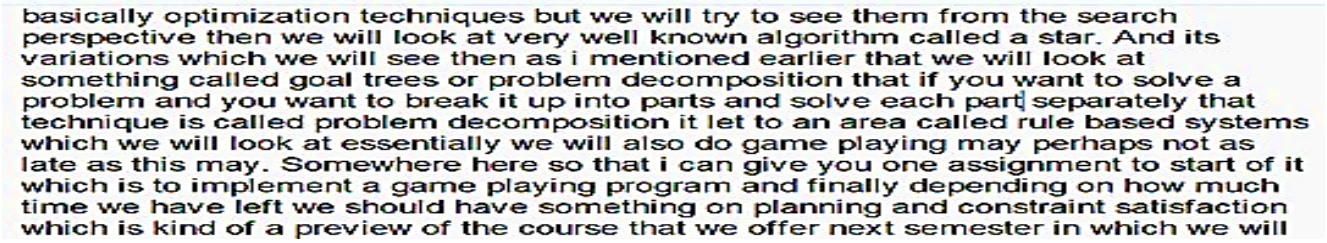
basically optimization techniques but we will try to see them from the search perspective then we will look at very well known algorithm called a star. And its variations which we will see then as i mentioned earlier that we will look at something called goal trees or problem decomposition that if you want to solve a problem and you want to break it up into parts and solve each part separately that technique is called problem decomposition it let to an area called rule based systems which we will look at essentially we will also do game playing may perhaps not as late as this may. Somewhere here so that i can give you one assignment to start of it which is to implement a game playing program and finally depending on how much time we have left we should have something on planning and constraint satisfaction which is kind of a preview of the course that we offer next semester in which we will

Fig. 3 Final Output

two outputs, we aim to assess the effectiveness of our refined and filtered approach in producing transcripts that are more accurate and contextually coherent. The results obtained provide valuable insights into the potential of our framework for advancing video summarization techniques through NLP.

Input Video URL https://www.youtube.com/watch?v=XCPZBD9lbVo&list=PLbMVogVj5nJQu5qwm-HmJgjmeGhsErvXD

Time Original Video 56:02 min Proposed Framework Input Time 29:00 min The Transcribe Generated by Whisper Model (Default)has been shown in Fig 2 The final Filtered Output by Our Framework is depicted in Fig 3 The Metadata Produced during Conversion as shown in table 1

Table 1 Metadata Produced

$Orig_file$	$Num audio chunks$	$Chunk_len_sec$	$Input_dur_mins$	$Num chars$	$Word count'$
NPTEL AI Video	58	30	29	17849	3330

6. Conclusion

In conclusion, this research paper explored the application of NLP techniques, specif- ically the Whisper model and wav2vec, in the task of video to transcript generation. The goal was to develop a system that could automatically generate accurate and reliable transcriptions of spoken content in videos.

Through our experiments and analysis, we have demonstrated the effectiveness of these NLP techniques in tackling the challenging task of video transcription. The Whis- per model, with its ability to convert audio to text, provided a strong foundation for

our approach. By leveraging wav2vec, which captures contextual information from the audio data, we were able to enhance the accuracy and robustness of the transcription system. Our results indicate that the combination of these techniques yielded promis- ing outcomes. The system achieved high transcription accuracy, successfully capturing the spoken content in videos across different domains and speakers. This suggests the potential for broader applications in areas such as automated closed captioning, video indexing, and content analysis. While our research has shown encouraging results, it is important to acknowledge the limitations of the current approach. Challenges such as background noise, speaker variability, and complex linguistic structures can still pose difficulties in achieving perfect transcriptions. Further research is needed to address these challenges and improve the system's performance. Looking ahead, the findings from this study open up avenues for future research in video to transcript generation. Exploring advanced NLP models, incorporating visual information from videos, and incorporating domain-specific knowledge are some directions that can enhance the accuracy and robustness of the system. Additionally, evaluating the system's perfor- mance on large-scale datasets and real-world scenarios will provide valuable insights for practical implementation.

To summarize, this research paper has demonstrated the potential of NLP tech- niques, including the Whisper model and wav2vec, in video to transcript generation. While there are challenges to overcome, this work contributes to the growing body of knowledge in this area and lays the foundation for further advancements. The development of accurate and efficient video transcription systems can have significant implications for various industries, making video content more accessible, searchable, and analysable.

Declarations

- Funding I would like to express my gratitude for the absence of funding in the creation of this manuscript, submitted for consideration to the Springer journal. This work was solely supported by the author's dedication, commitment, and passion for the subject matter.
- Conflict of interest/Competing interests None
- Ethics approval and consent to participate This article does not contain any studies with human participants performed by any of the authors.

References

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal processing magazine **18**(1), 32–80 (2001)

[2] Jayanthi, S.M., Pruthi, D., Neubig, G.: Neuspell: A neural spelling correction toolkit. arXiv preprint arXiv:2010.11085 (2020)

[3] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 439–448 (2016). IEEE

[4] Nagaraj, P., Muneeswaran, V., Rohith, B., Vasanth, B.S., Reddy, G.V.V., Teja, A.K.: Automated youtube video transcription to summarized text using natural language processing. In: 2023 International Conference on Computer Communication and Informatics (ICCCI), pp.

1–6 (2023). IEEE

[5] Singh, P., Srivastava, R., Rana, K., Kumar, V.: A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowledge-Based Systems **229**, 107316 (2021)

[6] Mathur, A., Saxena, T., Krishnamurthi, R.: Generating subtitles automatically using audio extraction and speech recognition. In: 2015 IEEE International Con- ference on Computational Intelligence & Communication Technology, pp. 621–626 (2015). IEEE

[7] Thangthai, A., Thatphithakkul, S., Thangthai, K., Namsanit, A.: Tsync- 3miti: Audiovisual speech synthesis database from found data. In: 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co- ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 77–82 (2020). IEEE

[8] Chen, L.-W., Rudnicky, A.: Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE

[9] Kadlˇc´ık, M., H´ajek, A., Kieslich, J., Winiecki, R.: A whisper transformer for audio captioning trained with synthetic captions and transfer learning. arXiv preprint arXiv:2305.09690 (2023)

[10] Porwal, K., Srivastava, H., Gupta, R., Pratap Mall, S., Gupta, N.: Video tran- scription and summarization using nlp. Proceedings of the Advancement in Electronics & Communication Engineering (2022)

[11] Vybhavi, A.N.S.S., Saroja, L.V., Duvvuru, J., Bayana, J.: Video transcript sum- marizer. In: 2022 International Mobile and Embedded Technology Conference (MECON), pp. 461–465 (2022). IEEE

[12] AlMousa, M., Benlamri, R., Khoury, R.: Nlp-enriched automatic video segmen- tation. In: 2018 6th International Conference on Multimedia Computing and Systems (ICMCS), pp. 1–6 (2018). IEEE

[13] Emad, A., Bassel, F., Refaat, M., Abdelhamed, M., Shorim, N., AbdelRaouf, A.: Automatic ideo summarization with timestamps using natural language pro- cessing text fusion. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0060–0066 (2021). IEEE

[14] Qi, W., Gu, L., Jiang, H., Chen, X.-R., Zhang, H.-J.: Integrating visual, audio and text analysis for news video. In: Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101), vol. 3, pp. 520–523 (2000). IEEE