Semantic Enrichment of Video Content using NLP Transformer Networks

M.Hanumesh[a], K. Sankar Brahmachari[a], Dr. G. Anitha[b], Dr. Balachandra Pattanaik[c]

{mhanumesh07@gmail.com, ksankarbc17@gmail.com, g_anitha@ch.amrita.edu, balapk1971@gmail.com}

[a] *Computer Science and Engineering,Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India*

[b] *Asst Professor, Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India*

[c]*School of Electrical and ComputerEngineering,Wallaga University, Nekemte,Ethiopia, Africa*

Abstract:The issue of managing and extracting important insights from the huge number of information offered in these gatherings has grown increasingly prominent in today's workplace, where meetings are a critical component of cooperation and decision-making. This study addresses the problem by presenting a technique for Semantic Enrichment of Video Content Using NLP Transformer Networks. The key issue is the requirement for a streamlined way to rapidly discovering and obtaining critical meeting content.The study looks at two key methodologies: the Video to Text API approach, which involves a multi-step process that includes video ingestion, pre-processing, audio extraction, audio-to-text transcription, and post-processing; and the NLP Transformers approach, which uses state-of-the-art Transformer models like OpenAI's GPT-3 for highly precise audio-to-text transcription, enabling automatic closed captioning, content indexing, and enhanced accessibility. The findings highlight the usefulness of both techniques in improving meeting material accessibility and efficiency, resulting in more informative, accessible, and time-efficient professional interactions.

Keywords: WER(Word Error Rate),API(Application Program Interface), Transformers, Uni-Directional, Self-Attention, Auto Regressive, Masked Self Attention.

---

1. Introduction

Video-to-text technology, turns spoken language into text format. It has grown in popularity and importance in various of applications across many industries. This technology analyzes video or audio input and transcribes it into text using advanced algorithms and machine learning models, making it a handy tool for various uses. According to the World Health Organization, impaired hearing affects [1] more than 5% of the world's population. These individuals must communicate with others via various digital means, such as videoconference meetings. As a result, the participant's speech must be converted into text that can be read in real time.Accessibility services are one of the most common uses of audio-to-text technology. Real-time captioning can help people with hearing impairments during live events or conversations, as well as in the workplace. By making spoken words available to others who may have trouble hearing, video-to-text technology has opened up new avenues for inclusive communication.

In the corporate world, video-to-text technology is critical in transcription services, where it can rapidly and accurately turn meetings, interviews, and conference calls into written documentation. This not only saves time but also improves documentation and accessibility, making it easier to search, analyzes, and share information.True intelligence is derived from the concept [2] of natural perception of our surroundings and the ability to accept input in a manner comparable to how humans see things through stimulation. To that end, speech is the most prevalent method of human connection, and while voice recognition is extremely essential, it has yet to live up to its full potential, and much research is being conducted to optimize the benefits of this technology.

Humans will become more reliant on machines as technology advances. Speech-based communication is the simplest way for humans and robots to communicate [3] because speech is one of the most easy communication mediums. The proposed solution is an application that translates participant’s audio into text during a real-time videoconference meeting. There are some solutions that achieve this, but the vast majority of them are commercial and not open source.We propose a transformational methodology that addresses the critical demand for textual conversion of videoconference conversations. The transcription of such meetings' spoken content holds enormous promise for improving information accessibility, analysis, and overall communication efficiency.

- Establishing use of two different models, Video-to-Text API and NLP Transformers, in order to assess their efficacy in addressing the semantic enrichment of video content.
- Demonstrating efficiency in faithfully transcribing videos >10 min,highlighting the adaptability and resilience of the selected models.
- Using the metrics such as Word Error Rate (WER) and Character Error Rate (CER) to demonstrate the superior performance of NLP Transformers.

---

2. Literature survey

Anggrainiet et al. acknowledges earlier research on speech recognition in the context of robots, [4] most notably comparing Google Speech with the Google Speech API, assessing characteristics such as word error rate and translation speed. In addition, the publication cites another work that presented a hierarchical probabilistic model-based solution for activity recognition using a Smartphone's accelerometer. In the paper's studies, it report a recognition rate of roughly 80% for speech-impaired individuals using their built speech recognition application, compared to a 100% recognition rate for normal speech. Furthermore, for speech-impaired voices, the recognition rate for spoken numerals from 1 to 10 ranged from 83.3% to 90%.

Vaishnavi K et al. [21] provided a comprehensive analysis of speech-to-text conversion techniques, encompassing Voice Recognition Module V3, Google Speech Recognition, PCM A-law codecs, CTC+Attention and End-to-end ST models. Their proposal provides a thorough overview of recent developments in the industry by examining a variety of methods and algorithms to accomplish necessary capabilities.

Rand Abdulwahid Albeer et al. [22] explored the use of TF-IDF for key term extraction in the automatic summarization of YouTube video transcriptions. The project uses natural language processing to automate summarising and give important information from long movies a succinct manner for researchers and students. The effectiveness of the suggested strategy is confirmed by evaluation using the Rouge method on the CNN-dailymail-master dataset.

Vetonet et al. compares Microsoft Speech API, Google Speech API, and Sphinx-4 [5] in a complete evaluation based on audio recordings and the performance metric Word Error Rate (WER). Google's API is particularly strong in acoustic and linguistic modelling. The TIMIT corpus is used in the study for acoustic-phonetic research and system development. For voice recognition, a Java utility connects with Sphinx-4, Microsoft API, and Google API, allowing comparisons with source texts. The article also goes over the evolution of Microsoft Speech API and its powerful speech platforms. In conclusion, Sphinx-4 had a WER of 37, Microsoft API had a WER of 18, and Google API outperformed with a WER of 9, proving its superior performance.

Gousiya Begumdelves into summarising YouTube videos [23] and how users might interact with it through a Chrome plugin. Furthermore, resources from the University of Edinburgh's XSum, Kapwing, Lofindo, and the Allen Institute for AI offer a variety of approaches, illustrating Begum's investigation within the larger field of YouTube video summarising strategies.

Takaaki et al.proposed a streaming transformer architecture that performs [6] with remarkable WER scores of 2.8% for "clean" and 7.3% for "other" test data from the LibriSpeech dataset, the transformer model performs exceptionally well in ASR. We use harmonic positional encodings and find that coupled CTC-attention decoding yields important benefits.

Venkateshet al. develops into speech-to-text translation using the [7] Hidden Markov Model (HMM) approach, which has shown superior precision when compared to other models. The authors describe speech-to-text conversion models in detail, stressing their proposed technique and HMM. Notably, the research investigates the integration of deep learning architectures such as MFCC, DWT, GMM-HMM, and DNN-HMM in the voice recognition training, testing, and assessment stages. In addition, the authors suggest a hybrid strategy that combines HMM and neural networks to improve speech-to-text conversion accuracy. This advancement holds promise, particularly for visually impaired folks, as it promises to make long texts easier to read.

Kurra et al. focuses on the creation of a comprehensive speech recognition system [8] customized for people with physical limitations, with capabilities spanning from text-to-voice and image-to-speech conversion to PDF-to-voice and speech-to-text conversion. The system's target users include those with various physical

disabilities, such as deafness, blindness, or differing abilities, as well as people who have difficulty typing or are unable to do so. The audiobook system is built with Python code and Visual Studio, resulting in a versatile and accessible solution. The system meets a variety of needs by incorporating four unique modules: text-to-voice, speech-to-text, PDF-to-speech, and picture-to-voice and text conversion.

Mamyrbayev et al. explores the utilization of Transformer models [9] for Kazakh speech recognition, transformer models and CTC were used to achieve a 3.7% character error rate on clean data. Features extraction is dependent on PLP and MFCC. DNN and HMM are used by the AM, and specific training is used to determine the neural network model targets. The study recommends LSTM or RNN for language modelling. Two hundred hours of "pure" speech and two hundred hours of impromptu phone speech from 380 native Kazakh speakers make up the speech corpus.

Siddiqueet et al. [10] review transformer applications in voice domains like as translation, synthesis, and voice recognition, highlighting their capacity to model long-term speech interactions while recognising limitations. The speech technology subfields are united by this effort.

G. Anitha et al. [11] examines the use of wearable sensors in health monitoring, highlighting the need of early aberrant activity detection in the elderly. Their suggested computer vision method prioritises privacy while achieving 82% accuracy on real-time video using a dynamic Bayesian network. The study advances ambient assisted living by improving age-related chronic disease support and emergency prevention.

G. Anitha et al.[12] emphasises the importance of vision-based methods for human fall detection by utilising a VEFED-DL model that combines GRU and MobileNet. The model, which uses a digital video camera and combines GTOA and SAE, performs better on datasets for many cameras and UR fall detection.

Lakomkinet al.[13] proposed that substitutes for speech datasets gathered by crowdsourcing because of restrictions on sample diversity and user engagement. Using multi-modal data sources such as Google Speech Commands, LibriSpeech, and TED lectures was mentioned. ASR systems were trained using data-cleaning techniques on YouTube subtitles. 150+ hours of YouTube voice data were gathered by the KT-voice-Crawler, which assessed the word mistake rate to be 3.5%. Using this data, errors on the Wall Street Journal dataset were lowered by 40%. 3.5% of words were assessed incorrectly by hand, mostly as a result of beginning or ending word errors.

Bogdanet al. [14] discusses the challengesin ASR employing cepstral analysis to attain a high word identification rate in languages with limited resources, such as Romanian. They use the Google Cloud Speech-to-Text API to transcribe Romanian e-learning content from YouTube. Their results show satisfactory accuracy with a WER score of 30.96, which compares favourably with other research of a similar nature.

Jacket et al. built on previous work using speech cues to enhance [15] understanding of online instructional videos. While rich captioning improved with visual and ASR features, the unique nature of instructional videos favored ASR-only models. The ATVideo model, combining visual and ASR-token-based characteristics, outperformed, as confirmed by statistical studies, showing its effectiveness in improving instructional video captioning through multimodal integration.

Yogeshet al.[16] carry out a thorough analysis of speech recognition systems, covering advancements, development processes, and multilingual applications. They do not, however, provide a thorough examination of linguistic diversity and system limitations, with a primary emphasis on English recognition. The study emphasises how important acoustic modelling is to the automation, usability, and effectiveness of ASR systems.

Jungyoonet al. [17]introduce CCVoice, a recording app for mobile that makes use of the Google Cloud Speech API. Using the programme, users record audio, and converted text files are arranged for web access with an emphasis on features that are easy to use and automated.

Yogita et al. [18] provide a multilingual speech-to-text system, but don't compare it to other approaches in a comparative analysis. They also don't deal with issues that arise in the real world, such speaking in a variety of voices and backgrounds. Evaluation in a variety of settings is crucial.

Nehaet al. [19] demonstrate a bidirectional Kalman filter-based real-time speech-to-text system that improves performance under noisy environments. It uses MFCC features, performs well compared to an HMM-based system, but doesn't have any specific database data. The system's word accuracy is 90%.

Harshaet al. Discuss[20]the relationship between overlapping speech, silence, and single-speaker speech in multi-party discussions is covered by Harsha et al. They provide a fresh approach to conversational analysis that improves speaker diarization and overlap identification.

Machine learning, deep learning, and bidirectional Kalman filters have all been shown to be quite good at transcribing standard spoken English for voice-to-text applications. However, difficulties arise when these algorithms are extended to convert text from meeting recordings and YouTube movies. The heterogeneous and unstructured character of content in these environments creates challenges for transcribing accuracy and reliability. Furthermore, in a meeting environment where numerous people talk at the same time, the increased risk of accidents reduces the efficacy of these strategies even further. The complex dynamics of overlapping speech patterns in such circumstances offer a significant constraint, making machine learning and deep learning less successful at reliably capturing and transcribing material.

---

3. Methodology

As we discussed above there have been different approaches which have been designed for this video file to text file now, we will be going through the famous approach video to text API on the one of the video or audio file.

1. VIDEO TO TEXT API APPROACH

The process of converting video to text using an API typically involves several key steps and components working in tandem. Initially, the video content is ingested into the system, either through URL links, direct file uploads, or real-time streaming. Following this, the API may engage in video pre-processing tasks, such as stabilization, noise reduction, and format conversion, ensuring that the video is in an ideal state for subsequent text extraction and analysis.In the meantime, the API extracts audio from the video. This is critical for digesting spoken stuff such as conversation or narrative.The method of extracting audio from a video file typically consists of three major parts. First, demultiplexing, also known as demuxing, separates the file's video and audio streams. The audio stream is next decoded, which involves transforming compressed audio data from formats such as AAC or MP3 back into a conventional audio format such as WAV. Finally, the decoded audio data is retrieved and saved as a standalone audio file or processed further to fulfil specific requirements.When using an API to transcribe audio to text, To improve audio quality, the audio data is first pre-processed, which includes noise reduction, audio normalization, and filtering. After that, the API extracts acoustic features from the audio, such as Mel-frequency cepstral coefficients (MFCCs) or spectrograms, which are used as inputs to the speech recognition model. This model, which was trained on a large dataset of audio and text, recognizes patterns in the audio that match to spoken words. To provide a time stamped transcription, the identified words are time-synchronized with the audio segments.Post-processing techniques such as punctuation and capitalization alterations, as well as language-specific fixes, may be used to increase readability and accuracy. Furthermore, the API may return confidence scores for each detected word, providing a measure of transcription reliability. These processes work together to improve the overall efficiency of the audio-to-text transcription process.

2. NLP TRNSFORMERS APPROACH

A transformational approach to video-to-text conversion is to use state-of-the-art Transformer models. These models, like as OpenAI's GPT-3, have demonstrated their ability to do natural language processing tasks and can be modified to transcribe spoken words in video material. It can translate spoken words into text with exceptional precision by putting the audio stream into the Transformer model. This method allows not just for the transcription of dialogues and speeches, but also for applications such as automatic closed captioning, content indexing, and increased accessibility for the deaf. The power of Transformers promises to make video-to-text conversion more accurate and efficient than ever before.

2.1 ENCODER

Connectionist Temporal Classification (CTC) is a technique used in automated speech recognition (ASR) with encoder-only transformer models such as Wav2Vec2, HuBERT, and M-CTC-T. The encoder part is employed exclusively in an encoder-only transformer model, which is the simplest type of transformer. This encoder takes an input sequence, often an audio waveform, and converts it into a series of hidden states known as output embeddings. An additional phase is incorporated in the CTC framework. This stage is responsible for providing

class label predictions by applying a linear mapping to the series of concealed states. These class designations are commonly represented by alphabetic characters (e.g., 'a', 'b', 'c', and so on). Using this method, the model can predict words in the target language with a relatively small classification head. This is due to the fact that the vocabulary it requires essentially consists of the 26 characters of the alphabet, as well as a few special tokens.

2.1.1 Bi-Directional

When processing each time step, the encoder takes into consideration both past and future context. In the case of audio data, this allows the model to evaluate not just the auditory properties preceding a specific point, but also those that follow. This bi-directional encoding aids in the acquisition of contextual information from the full audio sequence, which is critical for successful speech recognition.

2.1.2 Self-Attention

Self-awareness is a critical component of transformer models. When computing embeddings in the context of the encoder, self-attention allows the model to weigh the importance of different parts in the input sequence. When constructing each embedding in the bi-directional encoder, self-attention is used to consider the full series of hidden states. This means that each embedding is influenced not only by the individual input at that time step but also by relevant information across the sequence. This process of attention assists the model in learning complicated dependencies and relationships within the auditory data.

So, in the context of an encoder-only transformer model for ASR with CTC:

- The bi-directional encoding enables the model to consider both past and future context while processing audio data.
- The self-attention mechanism assists the model in attending to and weighing key information from the full sequence, allowing it to capture important patterns and dependencies for speech recognition.

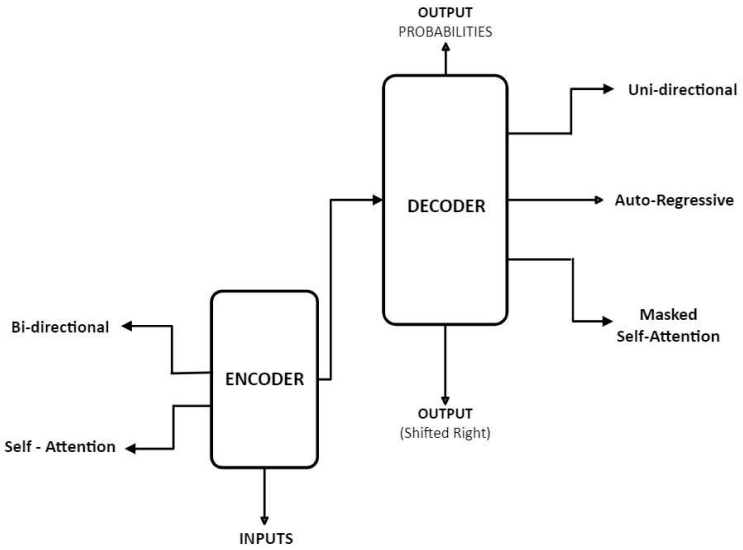

Fig 1. Model Architecture

2.2 DECODER

While the encoder processes incoming audio data to form embeddings or hidden states, the decoder's primary function is to generate transcriptions or recognized text from these embeddings.

2.2.1 Uni-Directional

In contrast to the encoder, which uses bi-directional processing to account both past and future context while constructing embeddings, the decoder is often uni-directional. The decoder processes the embeddings and generates transcriptions sequentially from left to right (start to finish) in unidirectional decoding. This is appropriate for activities like ASR, where we typically transcribe audio from beginning to end.

2.2.2 *Auto-Regressive*

The ASR decoder relies heavily on auto-regressive decoding. The decoder is auto-regressive, which implies that it generates transcriptions one token at a time, with each prediction influencing subsequent predictions. It considers previously generated tokens, making the production process dependent on the history of predictions. This is critical for preserving context and coherence in the transcribing.

2.2.3 *Masked Self-Attention*

Masked self-attention is a form of self-attention mechanism seen in decoders. The decoder uses self-attention to consider the embeddings of previously created tokens when creating transcriptions in an auto-regressive way. However, it must disguise future tokens in order to prevent "cheating" by attending to tokens that have not yet been generated. This masking ensures that each token forecast is based solely on relevant historical data and not on future data.

---

4. Results

We assessed three transcription models by analyzing Character Error Rate (CER) and Word Error Rate (WER). These metrics are crucial for evaluating the accuracy and quality of Automatic Speech Recognition (ASR) system transcriptions. Our analysis included both audio and video files as input sources, offering a comprehensive evaluation of the models.

Table 1. Model results on video file

Model	CER	WER
Transformer	0.15	1.2
CTC	0.17	1.8
Google API	0.25	2.3

Table 2. Model results on audio file

Model	CER	WER
Transformer	0.26	2.5
CTC	0.31	3.2
Google API	1.65	4.2

We observed kept track of the Word Error Rate and Character Error Rate for every model in our analysis of the data from various situations and datasets. Character and word-level accuracy precision was consistently correlated with lower CER and WER values. These results highlight how well the transcription models capture the subtleties of spoken language, indicating their applicability for a variety of uses.

---

5. Conclusion

Our research on video-to-text conversion APIs, specifically the Google Cloud API and the Transformers model, demonstrates the immense advantage of the Transformers approach. The Transformers model, which is known for its ASR capabilities, consistently supplanted the Google Cloud API in transcription quality, CER and WER. This study underlines the need of using advanced models that are suited to specific requirements in video-to-text conversion applications, strongly supporting for the use of Transformers models. As we analyze the video's content, our goal is to extract the important ideas and information offered by the speaker. Our summary approach will entail extensively scrutinizing the text created by the movie, allowing us to identify key concepts and noteworthy observations. By doing so, we hope to provide you with a streamlined and informed comprehension of the subject by delivering a condensed version that highlights the essential concepts of the film

---

6. REFERENCES

[1]. Sally, Eltenahy., Nihal, Fayez., M., Obayya., Fahmi, Khalifa. (2021). Conversion of Videoconference Speech into Text based on WebRTC and Web Speech APIs. doi: 10.1109/ITC-EGYPT52936.2021.9513968

[2]. Md., Tahsin, Tausif., Sayontan, Chowdhury., Md., Shiplu, Hawlader., Md., Hasanuzzaman., Hasnain, Heickal. (2018). Deep Learning Based Bangla Speech-to-Text Conversion. doi: 10.1109/CSII.2018.00016

[3]. Tanmay, Bhowmik., Aman, Rai., Soumya, Pandey., Prudhviraj, Boddu., Nilay, Patel., Veni, S., Vignatha, Manchala. (2021). A novel approach towards voice-based video content search. doi: 10.1109/I2CT51068.2021.9417

[4]. Anggraini, Nenny & Kuniawan, Angga & Wardhani, Luh & Hakiem, Nashrul. (2018). Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API. Telkomnika (Telecommunication Computing Electronics and Control). 16. 2733-2739. 10.12928/TELKOMNIKA.v16i6.9638.

[5]. Veton, Kepuska. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). International Journal of Engineering Research and Applications, doi: 10.9790/9622-0703022024

[6]. Niko, Moritz., Takaaki, Hori., Jonathan, Le. (2020). Streaming Automatic Speech Recognition with the Transformer Model. doi: 10.1109/ICASSP40776.2020.9054476

[7]. A., Elakkiya., Kumar, J, Surya., Konduru, Venkatesh., S., Aakash. (2022). Implementation of Speech to Text Conversion Using Hidden Markov Model. doi: 10.1109/ICECA55336.2022.10009602

[8]. Kurra, Santhi, Sri., Chennupati, Mounika., Kolluru, Yamini. (2022). Audiobooks that converts Text, Image, PDF-Audio & Speech-Text : for physically challenged & improving fluency. doi: 10.1109/ICICT54344.2022.9850872

[9]. Mamyrbayev, Orken., Oralbekova, Dina., Alimhan, Keylan., Turdalykyzy, Tolganay., Othman, Mohamed. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Dental science reports, 12(1) doi: 10.1038/s41598-022-12260-y

[10]. Siddique, Latif., Aun, Zaidi., Heriberto, Cuayáhuitl., Fahad, Shamshad., Moazzam, Shoukat., J., Qadir. (2023). Transformers in Speech Processing: A Survey. arXiv.org, doi: 10.48550/arXiv.2303.11607

[11]. G., Anitha., S., Baghavathi, Priya. (2019). Posture based health monitoring and unusual behavior recognition system for elderly using dynamic Bayesian network. Cluster Computing, 22(6):13583-13590. doi: 10.1007/S10586-018-2010-9

[12]. (2022). Vision Based Real Time Monitoring System for Elderly Fall Event Detection Using Deep Learning. Computer Systems: Science & Engineering, 42(1):87-103. doi: 10.32604/csse.2022.020361

[13]. Egor, Lakomkin., Sven, Magg., Cornelius, Weber., Stefan, Wermter. (2018). KT-Speech-Crawler: Automatic Dataset Construction for Speech Recognition from YouTube Videos. doi: 10.18653/V1/D18-2016

[14]. Bogdan, Iancu. (2019). Evaluating Google Speech-to-Text API's Performance for Romanian E-Learning Resources. doi: 10.12948/ISSN14531305/23.1.2019.02

[15]. Jack, Hessel., Bo, Pang., Zhenhai, Zhu., Radu, Soricut. (2019). A Case Study on Combining ASR and Visual Features for Generating Instructional Video Captions. arXiv: Computation and Language,

[16]. Yogesh, Kumar., Navdeep, Singh. (2019). A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature Review. doi: 10.1109/ICACTM.2019.8776714

[17]. Jungyoon, Choi., Haeyoung, Gill., Soobin, Ou., Yoojeong, Song., Jongwoo, Lee. (2018). Design of Voice to Text Conversion and Management Program Based on Google Cloud Speech API. 1452-1453. doi: 10.1109/CSCI46756.2018.00286

[18]. Yogita, H., Ghadage., Sushama, Shelke. (2016). Speech to text conversion for multilingual languages. 0236-0240. doi: 10.1109/ICCSP.2016.7754130

[19]. Neha, Sharma., Shipra, Sardana. (2016). A real time speech to text conversion system using bidirectional Kalman filter in Matlab. doi: 10.1109/ICACCI.2016.7732406

[20]. Sree, Harsha, Yella., Hervé, Bourlard. (2014). Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations. IEEE Transactions on Audio, Speech, and Language Processing, doi: 10.1109/TASLP.2014.2346315

[21]. R., K., Vaishnavi, N, R., Vaishnavi, K. (2023). Speech to Text App Customized for Police Functioning in Different Languages. 1-4. doi: 10.1109/INCET57972.2023.10170687

[22]. Rand, Abdulwahid, Albeer., Huda, F., AL-Shahad., Hiba, J., Aleqabie., Noor, D., Al-Shakarchy. (2022). Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency. Indonesian Journal of Electrical Engineering and Computer Science, 26(3):1512-1512. doi: 10.11591/ijeecs.v26.i3.pp1512-1519

[23]. Gousiya, Begum. (2023). Youtube transcript summarizer. International Journal of Progressive Research in Engineering Management and Science, doi: 10.58257/ijprems31301