An Approach for Audio/Text Summary Generation from Webinars/Online Meetings

Nitesh Bharti
Department of CSE
SRM University-AP
Andhra Pradesh, India
nitesh_bharti@srmap.edu.in

Shahab Nadeem Hashmi
Department of CSE
SRM University-AP
Andhra Pradesh, India
shahab_nadeem@srmap.edu.in

V. M. Manikandan
Department of CSE
SRM University-AP
Andhra Pradesh, India
manikandan.v@srmap.edu.in

2021 13th International Conference on Computational Intelligence and Communication Networks (CICN) | 978-1-7281-7696-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CICN51697.2021.9574684

***Abstract*—Due to the coronavirus disease (COVID-19) pandemic, most of the public work is carrying out online. Universities all around the globe moved to online education, job interviews are mainly conducting online, many first-level health consultations are happening online, and companies hold periodic meetings entirely online. Google Meet, Microsoft Team, and other online meeting software applications are widely accessible on the market. In this work, we are addressing a topic that has a lot of practical applications. In this paper, we present a method that takes a recorded video as an input and generates a written and/or audio summary of the same as an output. The suggested method can also be used to generate lecture notes from lecture videos, meeting minutes, subtitles, or storyline production from entertainment videos, among several other things. The suggested system takes the video's audio track, which is then transformed to text. In addition, we created the text summary utilising text summarising algorithms. The system's users have the option of using the text summary or creating an audio output that matches the text summary. The proposed method is implemented in Python, and the proposed scheme is evaluated using short videos acquired from YouTube. Since there is no benchmark measure for evaluating the efficiency and there is no specific dataset available for the relevant study, the proposed method is manually validated on the downloaded video set.**

***Index Terms*—Text summary, Audio summary, Video to audio conversion, Audio to text conversion.**

I. Introduction

The pandemic situation arose the need for most of the works in public sectors to be virtual. Spanning everything from online classes, job interviews to health consulting. A plethora of recordings is generated from the platforms utilized for these every day. Videos play a humongous role in sectors like education, entertainment, business, etc. Numerous times it's observed that due to both language and time barriers, certain informative components of the videos are overlooked by the user. Additionally, due to a bad network, the recording has an anomalous voice that makes it difficult for the person to understand. In general, subtitles of the respective video will help people to understand the video contents in a better way. The subtitles also help to comprehend the video contents having unknown audio contents. It may be noted that there are millions of videos are already available on the World Wide Web (WWW) without subtitles. In this research, we have considered the problem of generating subtitles from the video. In this paper, we proposed a scheme that will automatically generate subtitles from the given video sequence. Initially, the subtitles will be generated in the English language, and based on the user-specific language we will take the help of Google translate to translate the subtitles and summarize the video to fetch the important information that can cover the dropped voice issues as well as time constraints and additionally enable the participants to understand better. Our proposed system can be widely used in online teachings to help generate notes of the lectures, generate minutes of the meeting in the corporate sector, etc. The proposed scheme can be utilized to generate the subtitles and summary of the webinars or online meetings by incorporating it with video conferencing tools.

The pandemic situation caused by Covid - 19, affected a great ratio of the working sector, in order to adapt to the new situation, most of the business, corporate, and non-corporate firms got affected. To keep in check the progress and ensure the smooth operation of the various sectors that were affected, most of them went online and worked from home. The regular work updates, assignments, and, most notably, the education section, are all taking their work online on a regular basis. On such a frequent basis, the majority of them undertake regular meetings, classes, and lectures. As a result, a large amount of data in video format is generated on a regular basis. Sometimes, due to network failure or other unavoidable reasons, we are unable to devote the necessary time to these online meetings or pay attention.

It is also tough to go through previous recordings because it consumes a large amount of data as well as time. So, in this scenario, if we can obtain the whole transcript of the meetings or the summary of the video that we are referring to, we will be able to keep up with the pace and make better use of our time. This approach of obtaining the transcript will give similar kind of information and importance to that of attending online meetings or class lectures. It can even be applied to various educational or entertaining videos to improve understanding of the information in a short period of time. It is also useful for future reference or in-depth study of the video.

With the initial onset of the COVID-19 pandemic, many employees had to learn almost overnight how to use video conferencing, and our findings suggest many appear to have merely muddled through. As per [1] many employees were unaware of social norms or meeting etiquette as well as

978-1-7281-7696-3/21/$31.00©2021 IEEE
DOI: 10.1109/CICN51697.2021.2

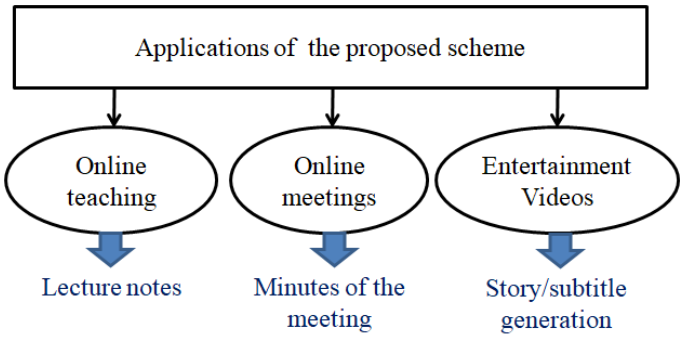

Fig. 1. Applications of the proposed scheme

suffering from lack of focus due to extensive work from home. Thereby originating a need for better and more understandable video conferencing features such as summary/minutes of the meeting, realtime captioning in the desired language which can be accessed as per the needs.

In our proposed method, we propose extracting the audio file from the given video file as an input in the first stage. We subsequently extract the video transcript from the whole audio file. This file can then be used to extract some useful information in order to obtain a full overview of the entire video clip. Thus we extract the summary of the video. Furthermore, we encourage translating the extracted transcript into any language and using it as a subtitle in the video to assist many people who are unfamiliar with the language spoken in it. We also propose text-to-voice conversion in case someone is more comfortable with the speech rather than the transcript. The primary goal of this paper is to extract the transcript from any video.

The proposed scheme will be useful in various domains, and they are listed below.

- Online teaching: In general, the online lectures will be recorded and it will be shared with the students for reference purpose. The proposed scheme can use to generate the lecture notes from the recorded lecture videos.
- Online meetings: Most of the meetings requires a minutes after it is done. During offline meeting, one dedicated person will be taking care of noting down the important points to prepare the minutes and finally it will be combined to prepare the final minutes. The scheme that we proposed in this work can be used to generate the minutes of the online meetings directly from the recorded meeting video.
- Entertainment: In entertainment domain, the proposed scheme can be used to generate story/subtitle from the entertainment videos.

The applications of the proposed scheme is graphically shown in Figure 1.

The following contents of this manuscript is organized as follows: in section II, we discussed about the related works in this domain, in section III, we detailed about the proposed scheme. In section III, we also described about all the modules used in the proposed work. Section IV demonstrates the experimental study and the results obtained.

II. Related Work

The exponential rise in the video based online teaching and virtual meetings has increased the importance of realtime audio to text based conversion for appropriate captioning and text summarization for getting the M.O.M i.e. Minutes of the meeting or summary. There have been a great deal of developments in the recent years over whether or not Text-to-Speech (TTS) or Spoken Language Processing (SLP) machines can synthesize natural synthetic speech from texts. This section briefly describes the methods that have been used for text summarization.

The authors have put forward a research into Semi-English sentences called Semi-English Language Recognition for Text Sequences (SELRTS) that people from different nations often use. Additionally, their research supplementally includes the writing rules or grammar for the better understandability of the language. The research was made using Linear Predictive Coding (LPC) algorithm which is an advanced Digital Signal Processing (DSP) filter used for synthesizing the input speech. The proposed method applies only to Finglish (a portmanteau term combining Farsi and English) and not on the Indian teaching methodology that may involve the cumulation of languages and also the text summarization that can have many uses cases/implementations in the real world scenarios.

Helal Uddin Mullah; Fidalizia Pyrtuh; L. Joyprakash Singh [3], proposed speech synthesis using HMM-based speech synthesis system (HTS) for the Indian English language. The trajectories of speech parameters for the proposed system were generated from the trained hidden Markov models (HMM). Additionally, the voice attributes such as the speaking style and emotions were easily detected in the system. The system lacks the text summarization that can have many uses cases/implementations in the real world scenarios.

A few related works and its details are given in Table I.

III. Proposed Scheme

The overview of the Proposed scheme is shown in Figure 2.

The modules in the proposed scheme are detailed below:

- Audio extraction module: This module will consider a video file (in any video format) as the input and it will extract only the audio track from the given video. The output of this module will be a waveform audio file with *.wav* extension.
- Audio to text conversion mdoule: This unit will consider the *.wav* file generated from the previous step and the corresponding text file will be generated. We have used built-in function available in Python to perform audio to text conversion. The summarization from audio files are difficult hence we used this module.
- Text summarization: This is one of the complicated task in the proposed scheme in which we have to consider a large text file as the input and the output should the summary details. This module expected to work in such

TABLE I
RELATED WORKS

Scheme	Research Methodology	Findings and Conclusions
Scheme [2]	The research was made using Linear Predictive Coding (LPC) algorithm which is an advanced Digital Signal Processing (DSP) filter.	The scheme allows conversion of Semi-English Language Recognition for Text Sequences (SELRTS) into english letters effectively.
Scheme [3]	The proposed scheme uses speech synthesis using HMM-based speech synthesis system (HTS) for the Indian English language.	The scheme effectively converts the audio into Indian English but the system lacks the text summarization that can have many uses cases/implementations in the real world scenarios.
Scheme [4]	The research was made using both abstractive and extractive text summarization.	The scheme uses statistical methods to demonstrate an extractive text summarization on a single document by utilizing both abstractive and extractive text summarization.
Scheme [5]	The proposed research was made using extractive text summarization.	The scheme used the extraction of the important sentence by the calculating word-frequency and phrase-frequency giving the useful measure of its significance.
Scheme [6]	The proposed scheme utilizes hierarchical structured self-attention mechanism that creates sentences and document embedding.	The research work asserts that the document embeddings delivers a better representation that enhances the text summarization and out-performs other models utilizing same dataset.
Scheme [7]	The research scheme utilizes multimodal abstractive text summarization and selective sentence routing.	The author effectively implements the Multimodal Heirarchical Selective Transformer modal(MHSF) and asserts that it out-performs other models.
Scheme [8]	The proposed scheme utilized the stop words from the text further filtering them by position features, overlapping and dependency word features to get the subject.	The scheme extracted the subject words from the urban complaint texts and had put forward a method to extract subject words.
Scheme [9]	The research was made using the metrics from text such as sentence position, sentence length, and average word frequency to evaluate the importance of casual sentences in the summary.	The research proposed a method for the extraction causal sentences from text documents and evaluation of the same based on some metrics to test how relevant they are to compose an extractive summary.
Scheme [10]	The research was conducted using the computer networks graph representation to capture the complete text structure where each sentence is modelled as node and relation as edge.	The proposed scheme reviews all the features that use metrics and concepts of complex networking for scoring sentences considering both quantitative and qualitative aspects.

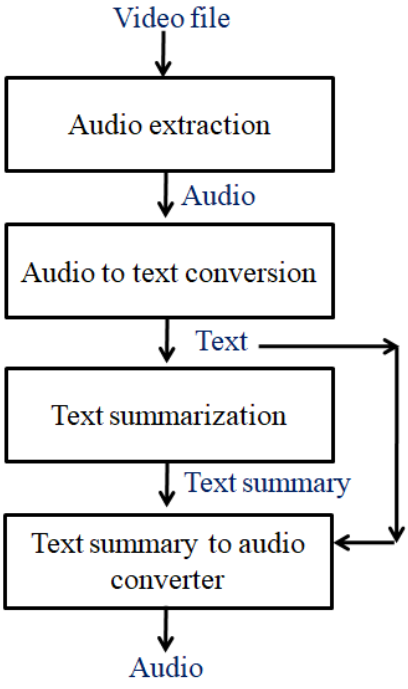

Fig. 2. Overview of the proposed scheme

a way that all the repeated text information and irrelevant text which is not related to the discussions should be removed.

- Text summary to audio converter: This module will use a text to audio converter for generating the audio output.

ALGORITHM 1: Proposed scheme for generating text summary from video

Input: A video file V of length T frames
Output: A text file F which stores the text summary corresponds to the video file

1 Extract the audio track A from the given video V.
2 Pass the audio file to a audio-to-text generator to get the text information corresponding to the audio file.
3 Apply text summarization technique on the generated text file to get the final text summary file F.
4 Return F

A. *Libraries used in the Proposed Scheme*

The major modules used in the proposed scheme are discussed here.

- The extraction of the transcript from the video file may be separated into three steps. The first step is to extract the audio clip from the supplied video file. Second, we turned the captured audio clip into a video transcript. Finally, a summary of the transcript file is created.
  Python and several built-in modules were used to implement the whole conversion. These modules are SpeechRecognition moviepy, pydub, spaCy, pyttsx3, pytextrank, and OS. The SpeechRecognition module sup-

ports a variety of recognition APIs. Google Speech API is one of them. MoviePy is a Python library that can read and write most common audio and video formats, including GIFs. The following two modules may be installed with the pip command, such as "pip install SpeechRecognition moviepy."

- Python has a package called 'pydub' that allows us to work with audio files, specifically .wav files. We may use this library to play, split, combine, and edit .wav audio files. We may also obtain information such as file length channels, volume increase or reduction, merging and splitting two or more audio files, and much more. This module may be installed using the command "pip install pydub."
- spaCy is a Python-based free and open-source Natural Language Processing (NLP) toolkit with a plethora of built-in features. It is used in NLP to process and analyze data. Unstructured textual data is generated on a huge scale, and it is critical to analyse and draw insights from it. In general, we represent data in a way that computers can understand. As a result, NLP plays an essential part in this. This module may be installed using the command "pip install spacy."
- Python's pyttsx3 module converts text to voice. It runs offline and is Python 2 and 3 compatible. To obtain a reference to a Pyttsx3, an application calls the pyttsx3.init() factory method. It's a simple tool that transforms text into speech. The pyttsx3 module supports two voices: one female and one male, both given by the "sapi5" Windows module. The command "pip install pyttsx3" will install this module. PyTextRank is a Python implementation of TextRank as a spaCy pipeline extension for graph-based natural language processing and knowledge graph activities. This covers the TextGraph algorithm family, which includes TextRank, PositionRank, and Biased TextRank. This library may be used for phrase extraction (obtaining the top-ranked phrases from a text document), low-cost extractive summarization of a text document, and assisting in the conversion of concepts from unstructured text to more structured representation.
- The Python OS module allows for the establishment of contact between the user and the operating system. It provides numerous helpful OS functions for performing OS-based activities and obtaining information about operating systems. The operating system is one of Python's basic utility modules. This module provides a portable mechanism to access operating system-specific functions.

B. Description of the functions used in Proposed Scheme

In order to reach our ultimate goal of obtaining the summary, we used a number of built-in and user-defined functions. The "moviepy" library contains the functions "VideoFileClip" and "write audiofile," which are used to read the video file and convert it to the desired audio file format. In our case, we deemed the file format to be '.wav.'

We also made leverage of a user-defined function called "get large audio transcription." The major objective of this function was to read an audio file of any length, divide it into small parts, and then proceed through the process of transcription one by one. Its execution time is entirely governed by the length of the audio file. We examined this using the length of silence or pause that most individuals use when speaking two or more phrases. The minimum length of silence considered is 500ms. If the silence is longer than 500ms, it divides the audio tape into smaller segments and extracts the transcript from there. The "recognize google" method of the "speech recognition" library assists in searching for suitable content from Google Speech to obtain the proper transcript. Finally, we aggregate the entire text into a single string and save it in the local device using the OS module.

The scheme's next step is to construct the summary from the extracted transcript. To do this, we used the "pytextrank" and the spacy libraries. These libraries extract some useful text depending on their rank and counts, and then use the notion of NLP to paraphrase the full sentence.

Finally, we added a text-to-audio option for the convenience of users who are referring to a long summary or the complete transcript. The "pyttsx3" module was used to accomplish this. It has the voice of two people: one with a female voice and one with a male voice. The accent featured in these voices is the American - accent. The input is any text file, either a transcript or a summary, and the output is speech through the use of the 'say()' and 'runAndWait()' methods. The "setProperty()" method of "pyttsx3" is used to input the id of the speaker we would like to use for audio. As a result, it completes the conversion of video to summary and adds the voice element for a better understanding of the transcript or summary of the entire video.

IV. Experimental Study and Result Analysis

The proposed scheme is experimented on 10 short video clips downloaded from YouTube. The results from the proposed scheme are validated manually since we do not have any standard video set for this kind of study and there is no standard metric to analyze the efficiency of proposed scheme.

The authors of this research analysed the video clips and presented their findings in Table II.

TABLE II
THE QUALITY OF AUDIO TO TEXT CONVERSION PROCESS

Video Number	Audio	Transcript	Summary
Video 1	3	3	2
Video 2	4	4	4
Video 3	5	5	4
Video 4	5	4	3
Video 5	5	4	5
Video 6	4	5	4
Video 7	4	5	5
Video 8	4	5	4
Video 9	5	5	4
Video 10	5	5	5

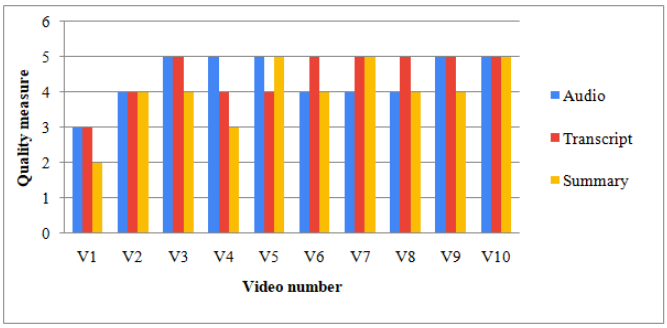

Fig. 3. The quality of audio, transcript and summary generated from the proposed scheme

On a scale of 1 to 5, the table describes the audio, transcript, and summary quality created by the proposed approach. The value 5 denotes a high-quality output, whereas the value 1 denotes a low-quality output. After a thorough examination of the video and audio quality, we discovered that if the audio quality is acceptable and the background noise in the audio is minimal, the transcript created is extremely good, and it also aids in effective text summary.

Figure 3 shows a bar graph that includes a comprehensive description of the various movies, their audio quality, a transcript, and a summary made using our proposed system. The audio and transcript quality of the videos V1, V2, and V3 is inferior to that of the others. This could be due to background noise or difficulties in distinguishing the speeches due to the speaker's accent.

Figure 4 and 5 shows a sample snapshot of the transcript and the summary. This demonstrates the amount of transcript that a typical video can generate, and then summarises it.

You have a group of people trying to introduce themselves. You ever get a new role within a team. Everything has a leader or directors and thanks for being smart sometimes Connaught. You forced to be intimate with complete strangers in asia maritime software suppliers of minutes. What have a good way to create a space that combine these two seemingly disturb accumulators. Intimate rupa people wear. Between occupation and head of something a bit more powerful working than the typical mandatory for events where I've never been. 52 42.0 in the military. Over 100 5.5 in a day with standard chartered chillara you enter question. Why you want date which is the saffron water on the project visit. Noorie married pregnant chilida. Naukari castrol w u l the phone is more how we would be to have tea represented to characterize the verses about understanding. Started this not traffic alerts on your forces will vary by the do that. Join us as we take away your sleep monologues. Email templates that person age race for the military alliances. Proof of incredible activities on them with incredible material reproduction ka use minimus passport office at no cost to no lights to reading at. Frog and princess on the language. And this show that either can be created any setting. Power of thinking room with complete strangers. And reminding us of human today in excess of expression is justice by both holes rifle and show their i can teknoparrot united amazon means. Socket pressure in the nose protecting our country.

Fig. 4. Transcript of the audio file

A group of people try to create a space where they can be intimate with complete strangers in asia maritime software suppliers of minutes. Join us as we take away your sleep by creating monologues that combine these two seemingly disturbingly disturbing accumulators. The aim is to remind us of the power of human beings and their ability to think for themselves even if they've never been there.

Fig. 5. Summary of the audio file

V. Conclusion

This paper proposes a scheme to generate the text summary and/or audio summary from a given video. The proposed scheme will have a lot of applications in the current scenario since most of the organizations are working online and most of the meetings are conducting online. The scheme introduced in this paper considers a video as the input and it will generate the text summary corresponding to that. The scheme can also generate the audio corresponding to the summary. The scheme can be utilized to generate the minutes of the meetings or it can be used to generate the lecture notes from a recorded lecture video. Our future work will be focused to make this proposed scheme as an add on to the well-known online meeting tools such as Google meet, Microsoft team, etc. The future work will explore the possibility of identifying the independent components in the audio sequence during a conversation and the person-wise identification of text summary can be generated.

References

[1] K. A. Karl, J. V. Peluchette, and N. Aghakhani, "Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly," *Small Group Research*, vol. 0, no. 0, p. 10464964211015286, 2021. [Online]. Available: https://doi.org/10.1177/10464964211015286

[2] M. S. Rafieee, S. Jafari, H. S. Ahmadi, and M. Jafari, "Considerations to spoken language recognition for text-to-speech applications," in *2011 UkSim 13th International Conference on Computer Modelling and Simulation*, 2011, pp. 304–309.

[3] H. U. Mullah, F. Pyrtuh, and L. J. Singh, "Development of an hmm-based speech synthesis system for indian english language," in *2015 International Symposium on Advanced Computing and Communication (ISACC)*, 2015, pp. 124–127.

[4] J. Madhuri and R. Ganesh Kumar, "Extractive text summarization using sentence ranking," in *2019 International Conference on Data Science and Communication (IconDSC)*, 2019, pp. 1–3.

[5] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Commun. ACM*, vol. 49, no. 9, p. 76–82, Sep. 2006. [Online]. Available: https://doi.org/10.1145/1151030.1151032

[6] K. Al-Sabahi, Z. Zuping, and M. Nadher, "A hierarchical structured self-attentive model for extractive document summarization (hssas)," *IEEE Access*, vol. 6, p. 24205–24212, 2018. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2018.2829199

[7] Z. Zhang, C. Zhang, Q. Zhao, and J. Li, "Abstractive sentence summarization with guidance of selective multimodal reference," 2021.

[8] Z. Dong, X. Lv, Z. Zhang, and X. Li, "Subject extraction method of urban complaint data," in *2017 IEEE International Conference on Big Knowledge (ICBK)*, 2017, pp. 179–182.

[9] C. Puente, A. Villa-Monte, L. Lanzarini, A. Sobrino, and J. A. Olivas, "Evaluation of causal sentences in automated summaries," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1–6.

[10] D. Cao and L. Xu, "Analysis of complex network methods for extractive automatic text summarization," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 2749–2756.