Semantic Aware Video Clipper Using Speech Recognition Toolkit

Adishwar Sharma[1[0000-0002-6458-8590]], Karanjot Singh[2[0000-0002-4833-550X]], Keshav Dubey[3[0000-0001-5171-5995]], Amit Kumar[4[0000-0002-3186-2530]], Prajakta Ugale[5]

[1,2,3,4,5] School of Computer Engineering and Technology, MIT Academy of Engineering, Pune, India
adishwarsharma@mitaoe.ac.in, kjsingh@mitaoe.ac.in, kkdubey@mitaoe.ac.in, agkumar@mitaoe.ac.in, pvugale@mitaoe.ac.in

Abstract. The Internet learning framework was presented as not a single one of us could meet up like previously, during the pandemic. Most understudies go to their web-based classes on portals like Google Meet, Microsoft Groups, Zoom. Students having some doubt regarding a topic, or they somehow missed the lesson, to cover the classes they can go through the recordings, but in these recordings, students are only interested in watching the content where the teacher was teaching a particular concept/topic, but the video also has a lot of the content which is not so useful, i.e., unrelated to the content being taught and also want to skip the part where nothing is happening. So, to eliminate these problems, we propose Semantic Aware Video Clipper (SAVC), a software to clip the irrelevant stuff and the part where no activity is there (silence) and make these videos/recordings relevant to the topic and hence save the priceless time of users, with respect to learning. SAVC will automatically cut some video fragments and then join these fragments together and this entire process will be completely automatic, without human intervention.

Keywords: Video Clipper, Semantic, Automatic, Silence.

1 Introduction

Please This pandemic has been a curse and blessing to us all. Let's keep the negativity aside and let's discuss about the blessing. As a blessing, it introduced online teaching system in our Indian education system, which was not at all entertained beforehand, new change for education System. As nowadays, all institutions are using platforms like teams, zoom for the online teaching, but after missing a lecture, students need to go through a recording to get along with the class. But the recording contains various things such as the part where no activity is there (silence) and the starting time of the class where teacher is waiting for the students to join, attendance, which is not relevant to the topic and watching all this will be time consuming.

Many advance video editing techniques are present to eliminate this problem. But video editing itself, is always time-consuming. Removing unnecessary video fragments is not a difficult, but time-consuming task. One must thoroughly review the

video (possibly multiple times!), select all necessary fragments, join them, and then render the video for an extended period. An hour-long video usually takes more than three hours to edit. So, SAVC (semantic aware video clipper) proposes, to trim the video automatically where no activity is taking place (silence), hence will help the students in saving their precious time while watching the recordings. In our future endeavors, we will be working on to trim stuff irrelevant to the topic being taught and make the video to the point.

2 Literature Survey

Sergey Podlsenyy et. al. [1] have proposed new features based on the state vector to dramatically improve the whole quality of the videos. It also provides an automated way of editing video footage from multiple cameras to create a meaning story of events present in any video. The steps involved in selecting the most valuable parts from the video in terms of visual as well as audio quality. The importance of the action planned was about cutting the footage into a meaningful story that would be interesting to watch. The extracts being used here was from ImageNet-trained convolutional neural network.

Tushar Sahoo and Sabyasachi Patra [2] have suggested a technique for composite silence removal under short time energy and statistical methods. The performance of the proposed algorithms was good as compared to STE and the statistical method. The proposed algorithm was applied in the first stage of speaker identification system and a 20% silence removal is identified after the final comparison.

Takuya Furukawa and Hironobu Fujiyo [3] have discussed a technique for editing personal videos and sensor information automatically (i.e., without human convention). The proposed technique uses CRIM (continuous rank increase measure and MC (motion correlation) values which generally capture any motion-change, acceleration, etc. from sensors to cut the scenes include object movement. Higher quality results were obtained using the suggested method.

Goutam Saha, S.S. Sandipan Chakroborty and Suman Senapati [4] have proposed a new silence removal and endpoint detection algorithm for speech-recognition applications. Their proposed algorithm used Probability Density Function (PDF) of the background noise and a linear pattern classifier for identifying silence part of the video. The proposed method also performs better end point detection and the silence removal than ZCR and STE methods.

A. Adjila, M. Ahfir and D. Ziadi [5] have suggested a method for detection and removal of silence from the audio. This method working is based on the continuous average of the speech signals. The technique used was beneficial to highlight the overall performance and, it's accuracy. But the audio from only three languages including English, Arabic and French performed better in noisy areas.

Dharmik Timbadia1 and Hardik Shah [6] have developed a straightforward computation for making the tests of the first audio records. Their suggested calculation gener-

alizes the clear edges and clamour by changing over sign. They have utilized MATLAB to generate substantial results.

Frederico Pereira et. al. [7] have presented a web-based speed-to-term recognition approach for providing an easier interaction with baseline functionalities by using a blend of techniques like Voice Activity Detection (VAD), Automatic Speech Recognition (ASR) and NLP (Natural Language Processing).

Prerana Das, Kakali Acharjee, Pranab Das and Vijay Prasad [8] have designed and implemented a voice recognition system for speech to text conversion. Their proposed system has two main parts i.e., first one for processing acoustic signal and the second one for interpreting. They have used the Hidden Markov Model (HMM) for building letters.

Babu Pandipati and Dr. R.Praveen Sam [9] have researched and evaluated the methods used in STT conversions. The method built on the interactive voice response was purposed. They explored various methods speech-to-text conversion for utilizing it in an e-mail system entirely based on voice. They have also suggested a model which uses both ANN and HMM techniques for STT conversion.

Brezeale et. al. [10] proposed various text, audio, and visual video classification techniques. They have suggested that only audio and visual feature extraction are used in various applications while being equally important. As a result, if they employ their combination, the results will be extremely accurate and precise. They have worked upon various automatic video classification techniques to assist viewers for finding best interest in viewing. They have also surveyed and discovered some features that are drawn from text, audio and visual modalities and furthermore investigated a wide range of feature and classification combinations.

Kranthi Kumar Rachavarapu et. al. [11] have addressed the major problems in data driven cinematography and video retargeting using gaze. An algorithm for content adaptation and automating the process of video creation was also purposed which edits with cut, pan and zoom operations by optimizing the path of a cropping window with the original video. They have also tried to preserve the crucial information. And they have focused on two algorithms i.e., one for efficient video content adaptation and the other for automating the whole process of video content creation. A novel approach was also presented by them in order to tackle the problems of efficient video content adaptation to optimally retarget the videos for varied displays with totally different aspect ratios and also by preserving important scene content.

Abdelkader Outtagarts et. al. [12] have presented a cloud-based collaboration and automatic video editing technique with an approach to automatic video editing from audio transcription to text using transcripts that are selected and automatically concatenated video sequences to create a new video. Their approach was based on extraction of keywords from audio-embedded in videos. They have proposed a keyword-based mashup modeling approach in which a video-editing testbed was designed and implemented. They have used keywords to automatically edit videos to create different mashups. It's been also stated that video editing is the process of selecting segments from a set of raw videos and then concatenating all of them by adding audio, effects and different transitions to create a new video or a type of mashup. Their proposed software wants to enable the automatic keyword-based video editing, deployed in the

cloud with a preview web-client. Different video which came from different video-sharing platforms or from webcams were stacked and also viewed on the fly. They have also suggested that the video editor is also a research testbed for the study of automatic video editing and summarization based on the statistical and textual data and it's metadata.
Edirlei Soares de Lima et. al. [13] have presented a real-time editing method for interactive storytelling. Their proposed method generates the most appropriate shot transitions, swaps video fragments to avoid jump cuts, and also creates ample looping scenes. Their research focuses on the concept of video-based storytelling to use pre-recorded videos with real actors. In their approach, automatic video editing was generally required for developing interactive narratives with the quality of feature films. But that was also their critical issue which was not fully addressed.
Favlo Vazquez et. al. [14] have suggested an idea to fix annoying silence in videos and all related errors. So, Python Jump cutter is used here to edit videos. Jump cutter is a program written in Python that automatically cuts silent parts of videos. The purpose method facilitates any post-recording work. It has one limitation in the form of very low video quality.
Joseph C. Tsai et. al. [15] have proposed a new motion-in-painting technique which allow the users to change the dynamic texture used in a video background for production of special effects. Motions estimations of global and local textures are used by them. Furthermore, they have used video blending techniques in conjunction.

In video editing, segments are selected from a set of raw videos and concatenated by adding audio, effects, and transitions to create a new video or mashup. Most of the video editing tools available are time-consuming and require specific knowledge. We discovered how automatic editing of multi-camera video footage works to create a coherent narrative of an event, where the components and training used are semantic data extraction. and model training for automatic editing. The steps to select the most valuable footage in terms of visual quality and action shot importance. Regarding video classification, different methods such as text, audio and video are used, but in different applications only visual and acoustic feature extraction are used, although they are equally important. We found that features are extracted from three modalities (text, audio, and image) and that a variety of feature and classification combinations were examined, while cropping windows within the original video while attempting to retain important information and follow cinematic principles.

3 System Design and Preliminaries

3.1 System Architecture

The below proposed diagram explains the overall architecture, Services and requests over the system. It acts as a blueprint that shows how to manage all the crucial service and requests over the system.

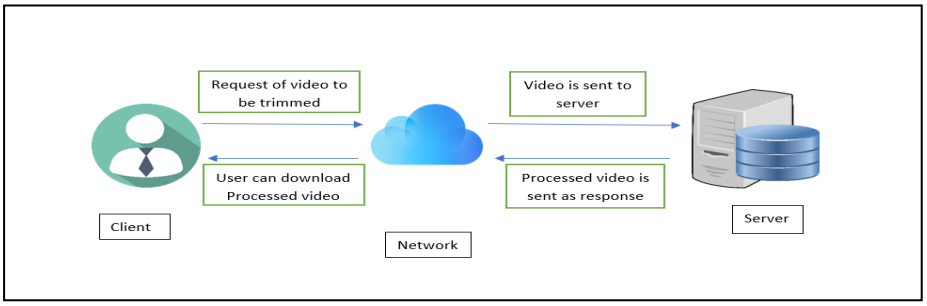

Fig. 1. Proposed System

The following diagram explains the overall flow of our project and shows the overall design of new systems and also provides a major overview of all the system components, key participants and the rest relationships of the proposed system. In this diagram, firstly, the user has to import the video in the given format i.e., avi, yuv, mkv, flv, wmv then the user will have gone to processing stage where the synchronization of video takes place and the video is automatically edited and once processing is completed, the file will get be ready to export.

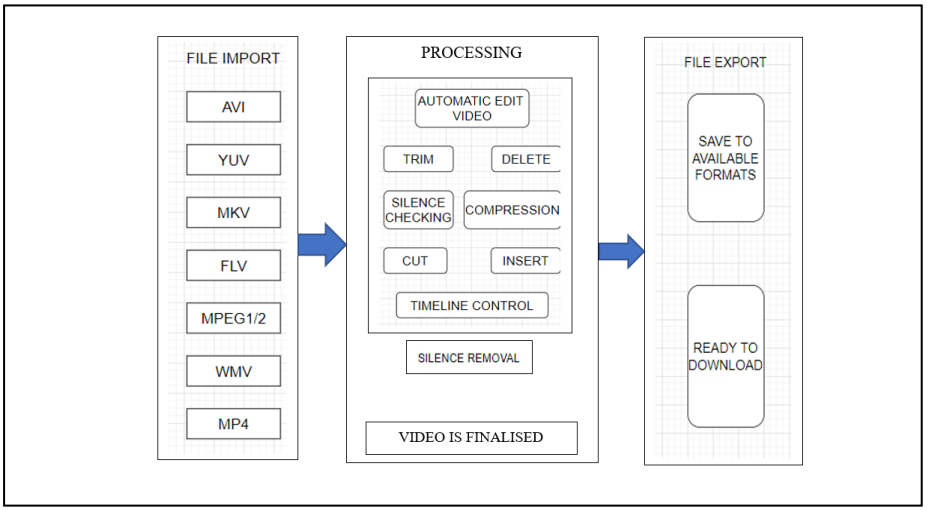

Fig. 2. Block Diagram

3.2 Preliminaries

Vosk API

Vosk is a speech recognition toolkit that supports more than 20 languages (e.g., English, German, Hindi, etc.) and dialects. It works offline and even on lightweight de-

vices like Raspberry Pi. Its portable models are only 50Mb each. However, there are also much larger models.

Silence Removal
This method cuts moments when silence (no activity happens) lasts longer than a certain threshold (for example, 2 seconds). This approach is fully automated and requires no human intervention during or after video recording. Just enter the video path and you will get the video without any silent moments.

Control Words Removal Algorithm
Control words are a set of commonly used words in a language. Control words are commonly used in text mining and natural language processing to remove words that are so commonly used that they carry very little useful information. This approach is being used to remove the part of the video which is not related much to the topic being taught in the lecture particularly for the recorded lectures from the college

4 Methodology

By using the software, user will drop/upload the video recording of the lecture afterwards it will process to backend in which it will iterate the length of the whole video and will store the intervals with no conversations/no audio (longer than 3 second's duration) will be detected using Machine Learning /Artificial Intelligence algorithm's and once the iteration is complete, it will clip out /trim the noted interval's using MoviePy library in python. And after the processing is completed, it will compress the video using High Efficiency Video Coding (HEVC) algorithm so that user can download the video while saving the data.

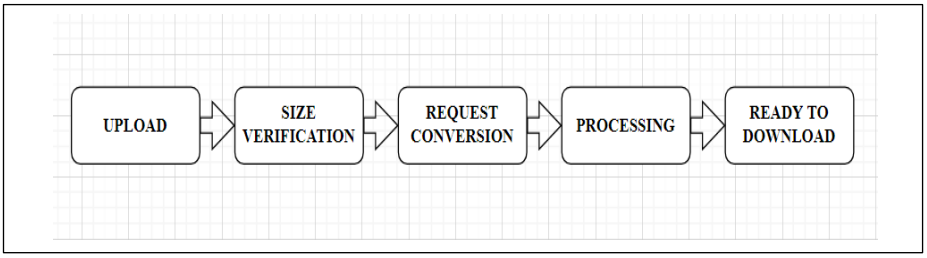

Fig. 3. General processing of the software

4.1 Techniques Used

The proposed system will automatically cut video into fragments in two ways:
A) Identifying long moments of silence, where no activity happens.

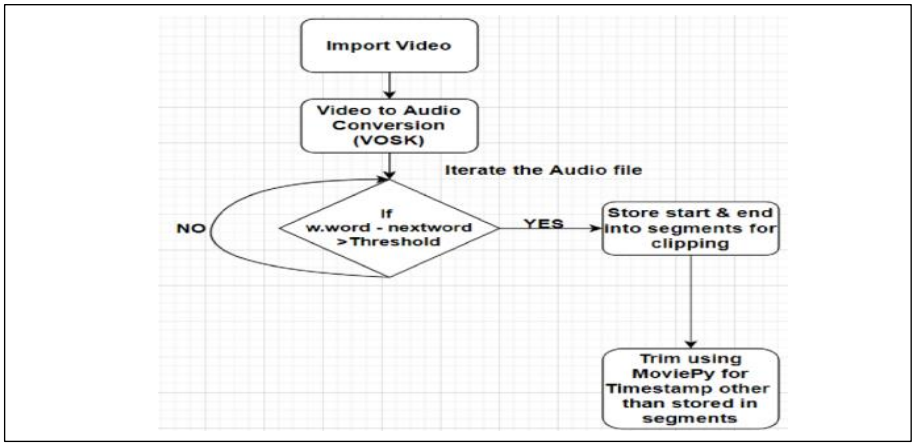

Fig. 4. Automatic Silence Removal

According to the proposed technique, the video will be imported first in the server, following to it, video-to-audio conversion will be done with the help of VOSK API. Now, it will iterate the audio file converted from the video, following by the conditions that if the time between the current word of a sentence and the first word of the adjacent sentence(next word) is greater than the threshold time set, then storing and clipping of the video segments, according to timestamps will be done and using MoviePy, trimming will be done according to the timestamp stored in the segments. Otherwise, it will keep iterate the audio file, until it finishes/ends.

B) Recognizing control word

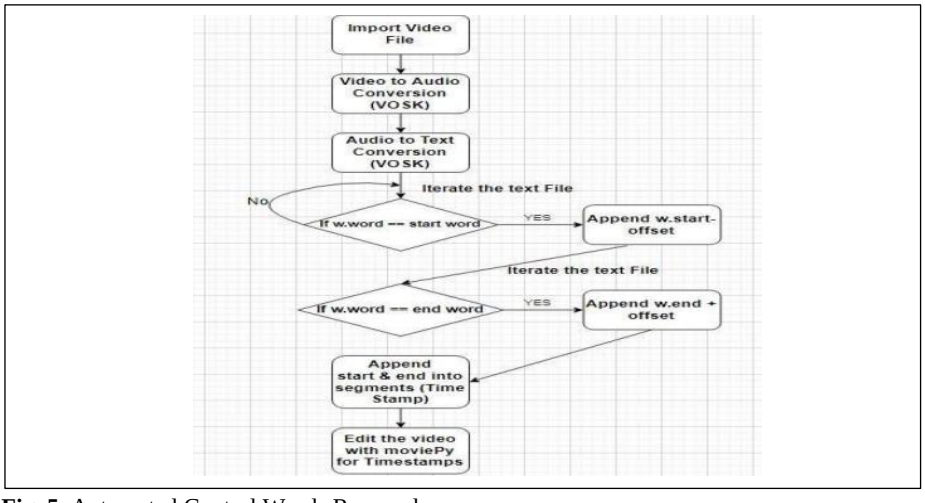

Fig. 5. Automated Control Words Removal

This method will be the future endeavor of the proposed system. With the proposed technique, the video will be imported first in the server, following to it, video-to-audio conversion followed by the audio-to-text conversion will be done with the help of VOSK API. Now, iteration of the text file will be done, converted from the audio, following by the conditions that if the start word of current sentence and end word of current sentence lies in the trained dataset of control words, it will append the updated timestamps in the list of timestamps by subtracting offset value from the respective timestamps, hence the timestamps to be included in the video would be ready by this time. Now, editing of I/P video will take place using Moviepy, and the processed video will be returned to the user as a result of request sent.

5 Results and Discussion

Video editing is always time-consuming. Removing unnecessary video fragments is not a difficult, but time-consuming task. One must thoroughly review the video (possibly multiple times!), select all necessary fragments, join them, and then render the video for an extended period. An hour-long video usually takes more than three hours to edit. The major targeted audience for the software is students, as to help them by saving their precious time while watching lecture recordings. As stated in the problem statement, the proposed software will automatically cut all video fragments and then join these fragments together and this entire process will be completely automatic, without human intervention. The suggested algorithm is very simple, as the various tasks involved are divided into 3 subtasks: ▪ Edit videos using moviepy library. ▪ Recognize control words/silence and their timestamps ▪ Connect these two components together This approach is fully automated and does not require any human intervention during or after video recording. One must specify the path to the video and get the processed video.

The below table shows the data collected during the testing phase of our project which was conducted throughout the month of May 2022 (on more than hundreds of videos), and these are some of the results, followed by the visual plotting of the values obtained. The testing was conducted by our group, to check various functionalities of the software being developed.

Table 1. Results among various processed videos

Original Video (IN SEC)	Silence (IN SEC)	Processed Video (IN SEC)
26	15	11
6	3	3
77	58	19
350	60	290
600	2	600
660	50	610

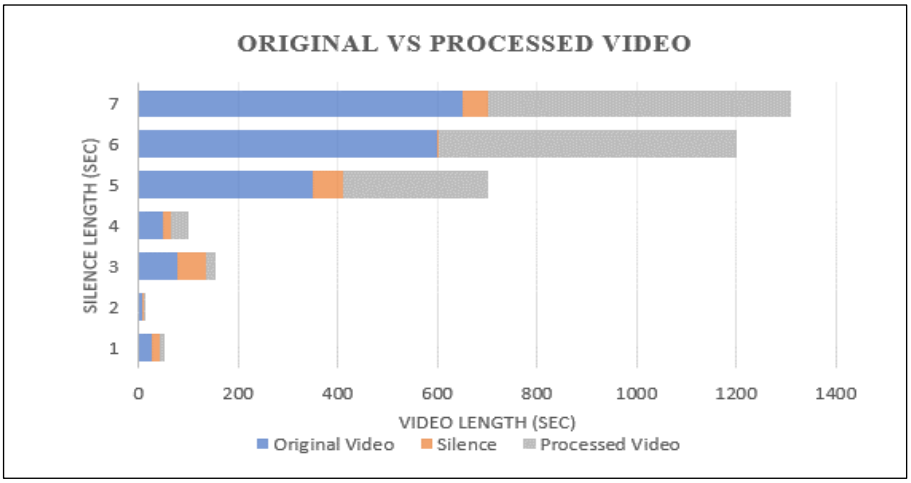

Fig. 6. Visual representation of the above table

6 Conclusion

Online learning has been a trend in the last decade as anyone can learn at their own place and whenever they like. Since pandemic, online learning has become one of the key factors in the field of education. Nowadays, majority of users are using platforms like teams, zoom for the online teaching, but if users miss a lecture, they need to go through a recording to get along with the class. So, the recording material should be short and up to the mark. The software focuses on using the available techniques to achieve many of these problems. The purposed software not only trims the silence part (no audio) of the recording but is also capable to reduce the memory size of the recording while retaining the original quality of the recording. Therefore, it is very feasible for users as they can directly access the relevant content and is considerable as a time-saving component for some extent. By examining all the above-mentioned features and benefits of the software, it has a lot of room for improvement, especially as India moves closer to becoming an educationally advanced country and more ideas and implementation in the field of online teaching and learning are needed. In our future endeavors, our goal will be to trim the stuff irrelevant to the topic being taught and make the video to the point.

The proposed software has some limitations as well, first being the time to process video. It totally depends on the length of video, the more the length of video, the more the time it takes to process the video. For processing a 10-minute video, it takes around 7 to 8 minutes. Another limitation is regarding internet connectivity. To process the video, good internet connectivity is required. Also, the time taken to process the video somehow depends on the internet connectivity as well.

7 Acknowledgement

We want to express our gratitude towards our respected faculty Dr. Vaishali Wangikar and Mrs. Vinodini Gupta for their constant encouragement and valuable guidance during the completion of this project work. We also want to express our gratitude towards respected School Dean Mrs. Ranjana Badre for her continuous encouragement. The success and final outcome of this project required a lot of guidance and assistance from many people, and we are extremely fortunate to have got this all along the completion of my project work. Whatever we have done is only due to such guidance. We would be failing our duty if we don't thank all the other staff and faculty members for their experienced advice and evergreen co-operation.

References

1. Podlesnyy, Sergey Y.: Automatic Video Editing, book release, 155-191 (2021).
2. Brezeale, Darin, Cook.: Automatic Video Classification: A Survey of the Literature, 416 – 430, (2008).
3. Kranthi Kumar Rachavarapu.: Towards Data-Driven Cinematography, (2019).
4. J. C. Tsai, T. K. Shih, K. Wattanachote, K. Li 2012.: Video Editing Using Motion Inpainting, 649-654 (2012).
5. A. Outtagarts, A. Mbodj.: A Cloud-Based Collaborative and Automatic Video Editor. In: 2012 IEEE International Symposium on Multimedia, pp. 380-381, IEEE, (2012).
6. E. S. d. Lima, B. Feijó, A. L. Furtado, A. Ciarlini, C. Pozzer.: Automatic Video Editing for Video-Based Interactive Storytelling, In: 2012 IEEE International Conference on Multimedia and Expo, pp. 806-811, IEEE, (2012).
7. Sahoo, Tushar, Patra, Sabyasachi.: Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification. In: 2014 International Journal of Image, Graphics and Signal Processing, pp. 27-35, (2014).
8. Saha, Goutam, Chakroborty, S.S., Senapati, Suman.: A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications (2005).
9. Furukawa, Takuya, Fujiyoshi, Hironobu.: A Cut Method for Cutting and Editing Personal Videos Using ST-patches and Sensor Information. In: The Journal of The Institute of Image Information and Television Engineers, pp. 93-100, (2012).
10. A. Adjila, M. Ahfir, D. Ziadi.: Silence Detection and Removal Method Based on the Continuous Average Energy of Speech Signal, In: 2021 International Conference on Information Systems and Advanced Technologies (ICISAT), pp. 1-5, (2021)
11. Pereira, Tiago, Matta, Arthur, Mayea, Carlos, Pereira, Frederico, Monroy, Nelson, Jorge, João, Rosa, Tiago, Salgado, Carlos, Lima, A., Machado, Ricardo-J, Magalhães, Luís, Adão, Telmo, Guevara Lopez, Miguel Angel, Garcia, Dibet.: A web-based Voice Interaction framework proposal for enhancing Information Systems user experience. Procedia Computer Science. 196.pp. 235-244, (2021).
12. Das, Prerana, Acharjee, Kakali, Das, Pranab, Prasad, Vijay.: VOICE RECOGNITION SYSTEM: SPEECH-TO-TEXT. Journal of Applied and Fundamental Sciences. 1, pp. 2395-5562, (2015).
13. Merabti, B., Christie, M., Bouatouch, K.: A Virtual Director Using Hidden Markov Models. In: Computer Graphics Forum, Wiley, 2015, 10.1111/cgf.12775. Hal-01244643.

14. Dharmik Timbadia, Hardik Shah.: Removing Silence and Noise using Audio Framing, pp. 118-120, (2021)
15. Vosk, https://alphacephei.com/vosk/adaptation
16. Moviepy, https://pypi.org/project/moviepy/
17. Video editing using python, https://towardsdatascience.com/make-your-own-video-editor-app-with-python-dash-moviepy-f0dd57c2b68e