Automatic Generation of Subtitle in Videos

Prachi Sharma[#1], Manasi Raj[#2], Pooja Jangam[#3], Sana Bhati[#4] ,Prof. Neelam Phadnis[#5]
[#]*B.E. Department of Computer Engineering, Mumbai University*
Mumbai, Maharashtra, India

Abstract

In this generation, people have been the witness of the emergence of videos and video plays a vital role. These videos entertain the human beings but sometimes it is hard for them to understand. Moreover, it is difficult for certain individuals who are deaf and have language barriers i.e. the difficulties in communication experienced by people or groups speaking different languages, or even dialects in some cases. Hence, there is a need to find a solution for such cases. There are many software utilities to generate subtitles for videos but needs manual participation of the user, which will make the work more tedious. Therefore, automated subtitle generation can be a possibility for such problem. This report will follow some standard by using speech recognition for generation of subtitle. There will be three stages namely audio extraction, speech recognition and subtitle generation respectively.

Keywords - *Audio Extraction, Speech Recognition, Subtitle Generation.*

I. INTRODUCTION

In this era, video is the most important aspect of entertainment. It is one of the most popular multimedia artefacts used in PC and Internet. As majority of the videos have sound in it and sound being the important phase of the multimedia some people may find it difficult to understand. People with hearing difficulties or gaps in spoken languages may need some help. Subtitles can help such human beings. It is essential to use subtitle for videos. Manual creation of subtitle is long and boring activity. So considering automatic subtitle generation seems to be a valid subject of research. In this report, it includes the user not downloading the subtitles from the internet and rather getting subtitle using automatic subtitle generation. This report can help eliminate the long procedure used in manual generation of subtitles and may give us an appropriate result. This can be achieved if we are aware of the content of the videos. The knowledge about the content of the videos comes from the metadata of the content. The metadata can be stored along with the videos as annotations. And as manual classification of videos does not give quality results, so will manual annotation of videos. This has given rise to the need of automatic and unsupervised classification and annotation of videos.

II. LITERATURE SURVEY REVIEW

We got this overview from some papers about automatic subtitle generation. There are many programming language available for the creation of subtitles.

A. Background Study

1. Video Selection and Suitable Language Program

A literature survey was needed to understand the functionality of automatic subtitle generation. We found out there are several methods in which we can implement our system. There are some positive analyses about the programming language. C++ provides speed, cross system package and well-tested packages. Whereas, Java offers intuitive syntax, portability in various operating system and trusted libraries. We found out that C# language is rather programmable than the other programming languages. Three distinct modules have been defined namely audio extraction, speech recognition and subtitle generation. The system should take video as an input and give subtitle as an output or the result with video being of a particular format. In audio extraction, it should take certain type of video format suitable for the speech recognition part. It should be capable enough to verify the input so that it can give the audio extraction feasibility. When it gets the proper input the audio is extracted and it is used in further stage i.e. speech recognition. Speech recognition is the important part of this thesis. It affects the performance and the result directly. It is the key part of the automatic subtitle generation. It gets the type of input (wav,avi,etc) and if it is the correct type of file then speech recognition is applied to it. It should be able to detect the silences so that it does not give continuous words as an output when it is not necessary. Subtitle generation gives the subtitle of the particular video. Time synchronization is the main part in subtitle generation. [1]

a. Audio Extraction

This module deals with audio extraction from video file. It generates audio file from a video, the embedded audio stream is taken from the video file

and then it is converted into audio file. Here we are trying to isolate the audio part from the video file. The video is distinguished on the basis of different aspects depending on the video's tone, depth, pitch, etc. it consists of four stages and they are start extraction, set properties, set audio frames in buffer, store audio in buffer and save the audio file.

b. Speech to Text Conversion

In speech extraction, we are going to use

Windows Media Player Format because of its simplicity. Firstly, .NET works efficiently with windows environment; it supports all types of format and it is easy to implement. This phase is primarily responsible for text creation. It involves speech recognition engine to break down the audio format to text form. It involves different types of methods for speech sound so that it can differentiate between multiple users or speakers in the video. Speech recognition is difficult when there is high vocabulary used or a word used which is similar in sounding. It also gives rise to synchronization problem. The text will probably not match with the time-stamp of the video.

c. Subtitle Creation

The analysis and design taken in consideration will give a module generating subtitle file of text format.

III. PROPOSED SYSTEM

We are presenting a system, which is broken down into three stages to give the final output as the subtitle for a specific input video.

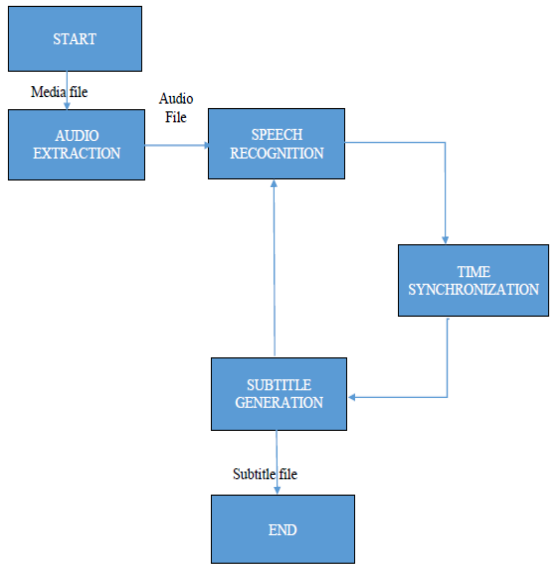


Fig 3.1: System Architecture

The proposed system to be implemented is described briefly and can be depicted as in the above diagram.

A. Audio Extraction

This part deals with extraction of audio from the video to deliver proper speech to text conversion. It generates an audio file from a video. This is accomplished by FFmpeg. FFmpeg is a free software which provides complete, cross-platform solution to record, convert and stream audio and video. The format of audio file can be of any type like wmv, flv, avi, etc. For more flexibility, convenience and efficiency APIs can be used for Audio Extraction. FFmpeg libraries are used to do most of our multimedia tasks quickly and easily say, audio compression, audio/video format conversion, extract images from a video and a lot more. Developers for transcoding, streaming and playing, can use it. It is very stable framework for translating of video to audio. Audio Extraction is divided into two parts: Speech extraction and Non Speech sound Extraction: For the implementation, part of audio extraction the Windows Media format SDK is used because of the following reasons: It works well with Windows Environment specially .NET framework and it supports almost all kind of video formats available in today's time.[5]

B. Speech Recognition

This segment is responsible for converting the audio file into textual format. It involves recognition of speech, which is useful for speech recognition processing task. It is a task of pattern recognition. The major steps involved are normalization, parameterization, comparison and decision. Each acoustic sequence of speech sample is classified as one of a set of linguistic category. If the utterance is the single word, we search for the best match among all the words in the vocabulary. Different methods are used for speech sounds, non-speech sounds and for multi-speaker system i.e. to distinguish between multiple speakers. In this phase, SAPI speech engine is used. SAPI (Speech Application Program Interface) is an application program interface (API) provided with the Microsoft Windows operating system that allows programmers to write programs that offer text to speech and speech recognition capabilities. The speech recognition feature is helpful for our system to convert the audio into appropriate subtitle. SAPI has some important components, which is useful for our thesis. There is one component known as Direct Speech Recognition, which is a low-level interface similar to voice command. The audio directly connects to the speech engine to give application more control and speed. The other component, which is important, is audio object. It tells the speech engine where to gets its audio. [3][5]

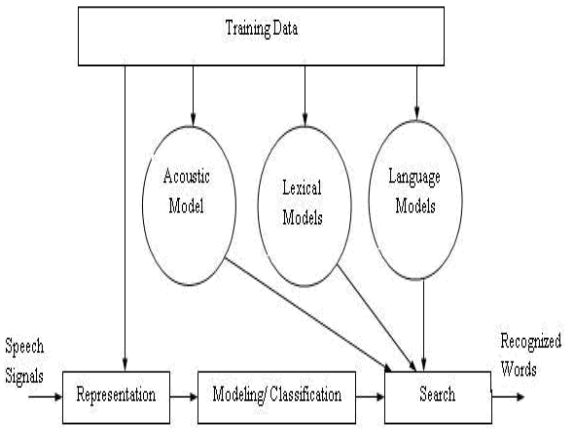


Fig 3.2: Components of Speech Recognition

The above figure shows the major component of speech recognition system. The digitized speech signal is first transformed into a set of useful measurements at a fixed rate, typically once every 10-20 msec.

These measurements are then used for searching the most likely word candidate, making use of constraints imposed by the acoustic, lexical and language model. Throughout this process the training data is used to determine the values of model parameters.

C. Subtitle Generation

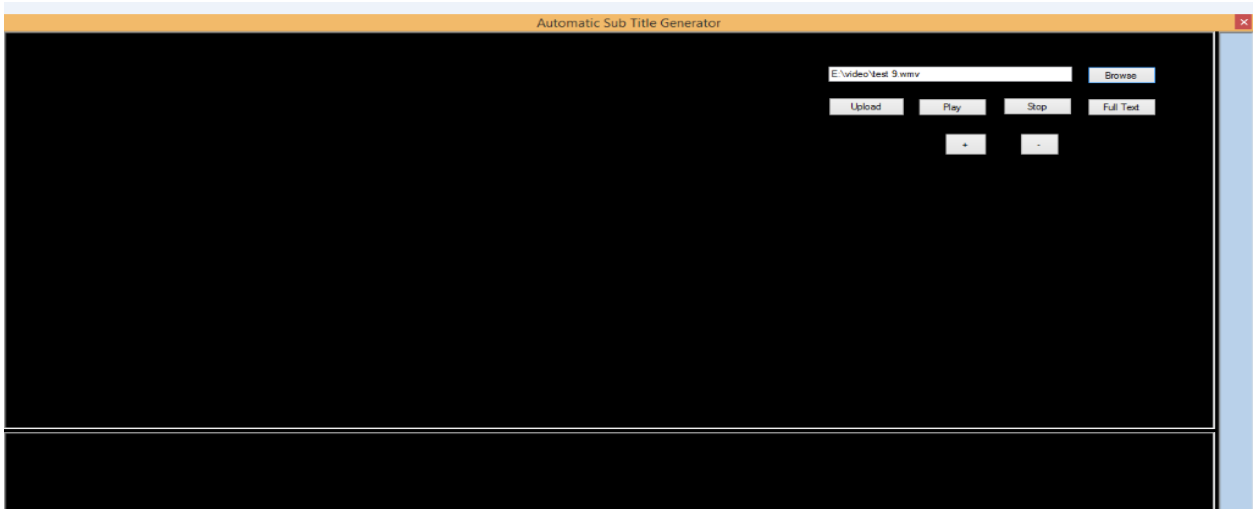
The subtitle generation routine is expected to get a list of words and their respective speech time frames from speech recognition module and then it produces a subtitle file. It also aims to create and write in a file in order to add multiple chunks of text corresponding to utterances limited by silences and their respective start and end times. Time synchronization considerations are of main importance. The file contains of the words that are spoken in audio file along with the time intervals in which it occurs in the video. This file is then further displayed with the video along with the time synchronization. However we had certain limitations in our system i.e. we are not able to define punctuations in our system since it involves more speech analysis. If the user wishes to check the entire subtitle of the respective video file then we are providing a pop-up window where user can see full text of that particular video. [2]

IV. IMPLEMENTATION OF THE SYSTEM

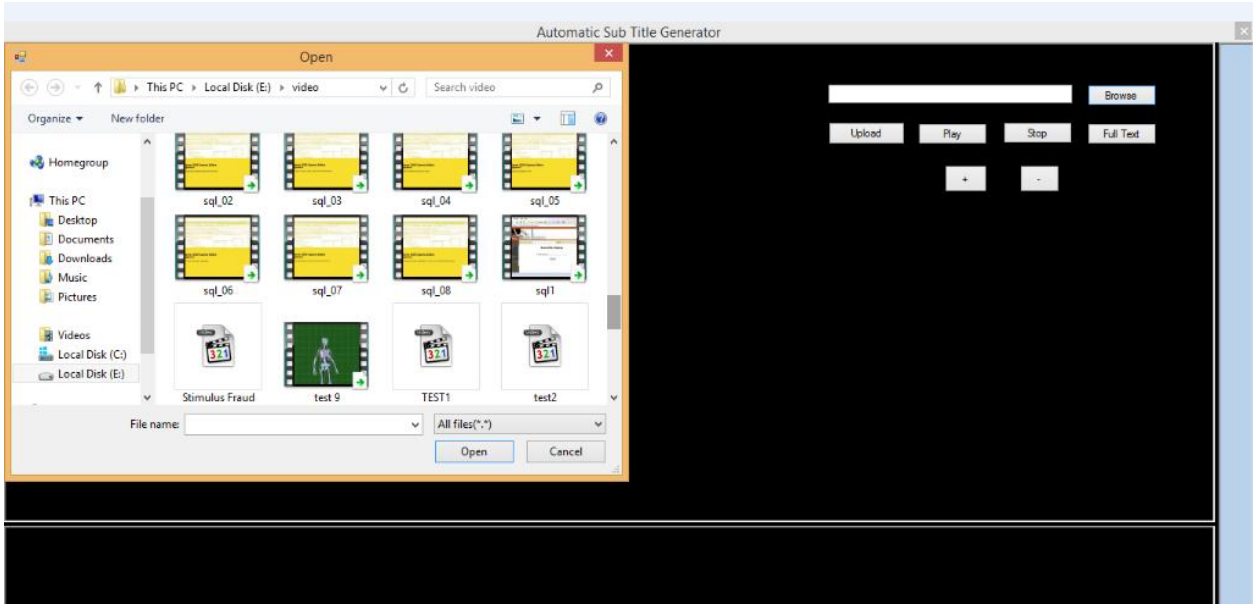
A. Graphical User Interface

This is our systems GUI, which will appear to the user. We, have provided different buttons in our system. Buttons like play, stop, + (To increase the volume), - (to decrease the volume).

B. Input Video File Page

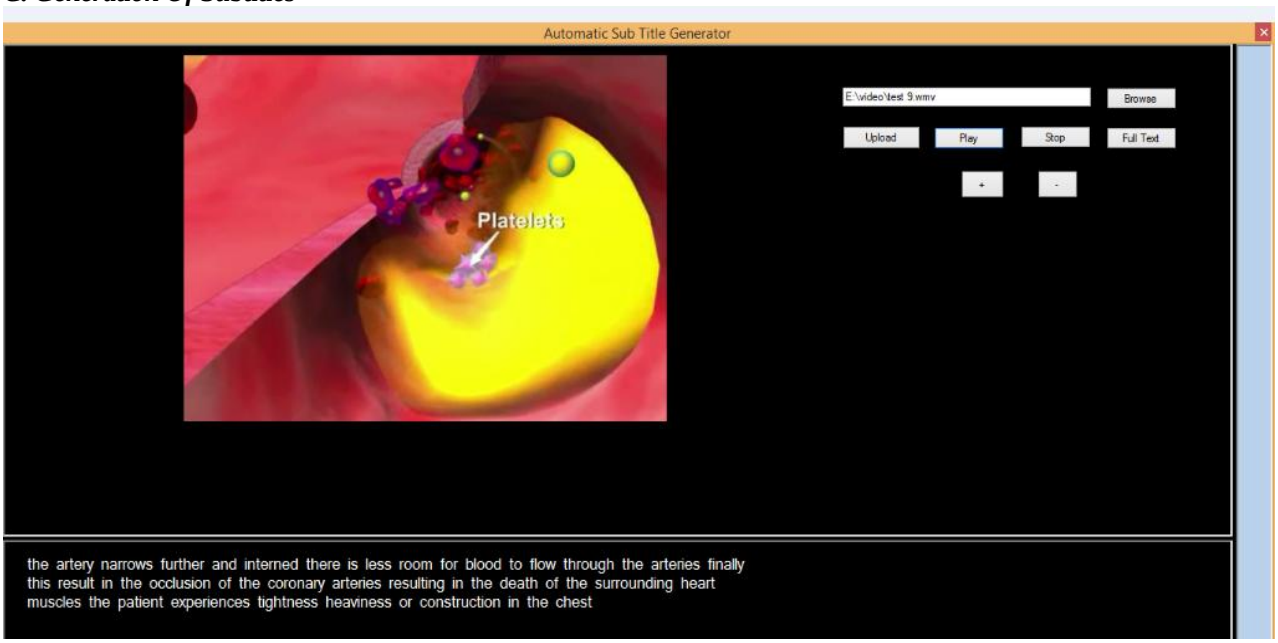


Here, the user can browse the video file that he wishes to upload.But the user can upload only those video files which are supported by windows media player .



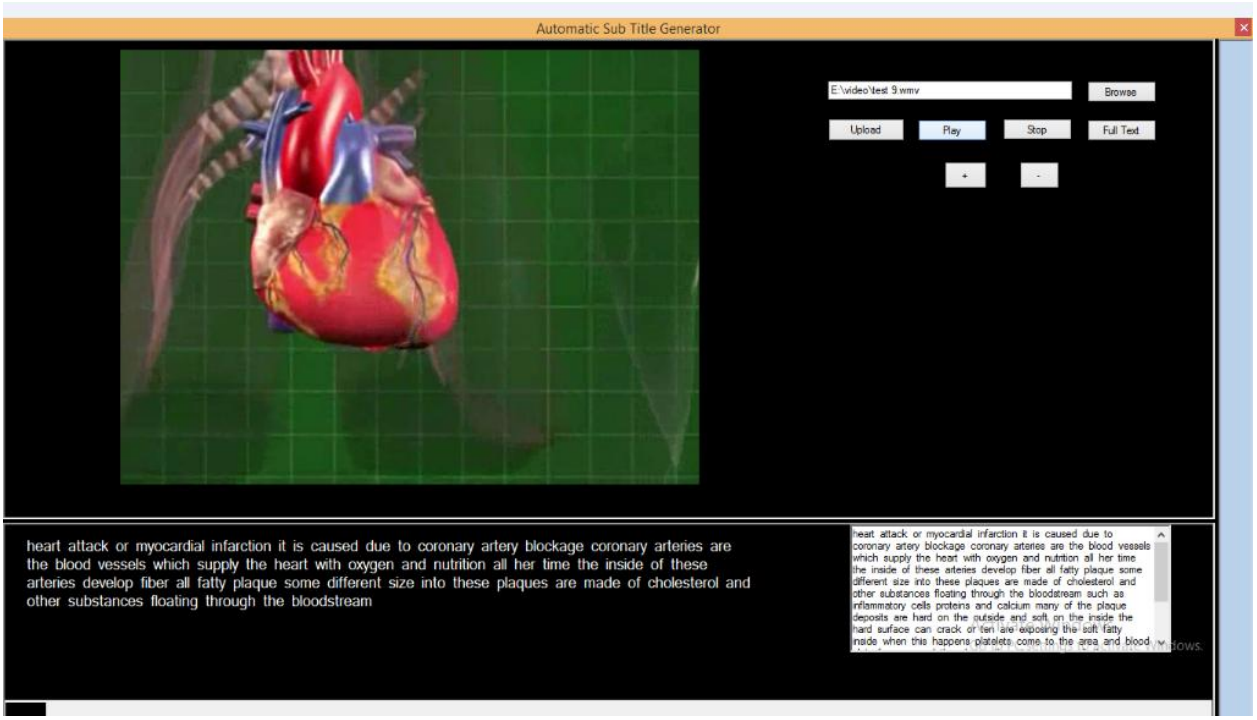
After choosing a suitable format .The user can upload that video for which he/she wants the subtitles, and our system will start processing to get the subtitles.

C. Generation Of Subtitles



Once the user has uploaded the video by clicking the upload button. Audio extraction ,Speech recognition and Subtitle generation process would be initiated. This will hardly take a minute .And after that subtitle file will be displayed to the user along with time synchronization.

D. Full Text



Apart from this if the user is not satisfied with the subtitles and if user want to see the full text then the user can simply click on Full text button and the entire text will be displayed as shown in the figure.The user can scrolll up or down as per users need.

V. CONCLUSION

We have implemented automatic generation of subtitle in videos in which three modules are included. Firstly, the Audio Extraction where the audio will be extracted into the audio format from the video that is given as an input by the user. Audio Extraction is done by using FFmpeg .Then the second phase is Speech Recognition where the words are extracted from the audio and results into the subtitle file which contains the text of the input video file. Subtitle is timely synchronized with the video. We have also provided an option where the user can check the entire subtitle of video in one go.

REFERENCE

[1] Boris Guenebaut, "Automatic Subtitle Generation for Sound in Videos", Tapro 02,University West, pp. 35, 2009.
[2] International Conference on Computational Intelligence Communication Technology 2015 IEEE Transaction for Generating Subtitles Automatically using Audio Extraction and Speech Recognition.
[3] Santosh K. Gaikwad, BhartiW. Gawali, Pravin Yannawar, "A Review on Speech Recognition", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.
[4] Akshay Jakhotiya, Ketan Kulkarni, Chinmay Inamdar, Bhushan Mahajan, Alka Londhe "Automatic Subtitle Generation for English Language Videos" SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – volume 2 issue 10 October 2015
[5] https://whatis.techtarget.com/definition/Speech-Application-Program-Interface-SAPI