An Approach Towards Generating Subtitles Automatically from Videos by Extracting Audio

Rizwan Sheikh
UG Student
Department of Computer Science and Engineering
P. R. Pote (Patil) College of Engineering Amravati, Maharashtra, India

Swapnil Suryajoshi
UG Student
Department of Computer Science and Engineering
P. R. Pote (Patil) College of Engineering Amravati, Maharashtra, India

Shivam Gupta
UG Student
Department of Computer Science and Engineering
P. R. Pote (Patil) College of Engineering Amravati, Maharashtra, India

Sushant Tayde
UG Student
Department of Computer Science and Engineering
P. R. Pote (Patil) College of Engineering Amravati, Maharashtra, India

M. S. Burange
Assistant Professor
Department of Computer Science and Engineering
P. R. Pote (Patil) College of Engineering Amravati, Maharashtra, India

Abstract

Videos are very important and helpful in our daily life for understanding and comprehend the information. With the help of videos user can gain the knowledge and spreads that knowledge to the other peoples. Hence, here it becomes important to make videos available to the people having auditory problems and even more for the people to remove the gaps of their native language. This all things are getting by using the subtitles for videos. There are several websites which are providing the subtitle files for the videos but not for all videos. Downloading subtitles from internet is a monotonous process. Hence to generate subtitles automatically with the help of software is the valid subject to research, this research paper resolves the problems mentioned above through the speech recognition technology. There are three models helps to generate subtitles for videos, Audio Extraction helps to extract audio from video and convert into .wav format for speech recognition process. Here 24% reduction rate has been achieved in the size of the video after the extraction. With the help of PocketSphinx speech recognition engine the .wav file get processed and the speech is get converted into text format and stored in .srt file for future use.

Keywords- Sphinx, Ffmpeg, Videos

I. INTRODUCTION

In today's world, the use of subtitles has become important for understanding of the video. Subtitling is essential for people who are deaf, who have reading and literacy problems, and can to those who are learning to read [5]. Subtitles provide information for individuals who have difficulty understanding the speech and auditory components of the visual. This leads to a valid subject of research in the field of automatic subtitle generation. Thus, this paper provides the users a major benefit of not downloading the subtitles from the internet instead generating them automatically.

At present, In Video LAN Client (VLC) media player the subtitles must be inserted first to the media player and then it is synchronized with the song. [2] The file inserted needs to be .srt file containing the time intervals of the text spoken. It does not accept .txt file to synchronize the subtitles as it only contains the lyrics of the song and not the time intervals of the spoken words. [2] On the other hand, YouTube accepts both the .txt file and the .srt file for synchronizing the text. Nonetheless, software generating subtitles without intervention of an individual using speech recognition have not been developed.

II. MECHANISM

System Design focuses on how to accomplish the objective of the system. Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Three distinct modules have been defined as, Audio Extraction, Speech Recognition and Subtitle Generation which take a video file as an input and generate subtitle file. Fig. 1 shows the interaction between various stages of the system. Following is the brief description of each stages of automatic subtitles generation process:

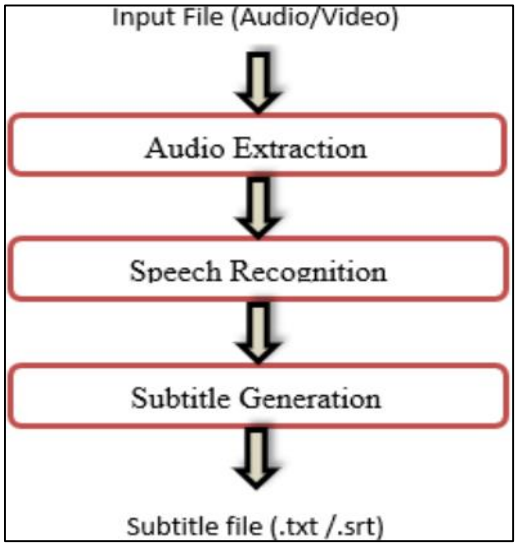

Fig. 1: Three stage process of automatic subtitle generator

A. Audio Extraction

This module aims to output an audio file from a media file. It takes as input a file URL and gives audio format of the output. The speech recognition engine requires a .wav audio file as input. Hence, it is necessary to convert the video into .wav format [2]. Ffmpeg is a complete, cross platform solution to convert and stream audio and video. [7] Audio extraction provides a way to extract audio from a media file. It specially uses ffmpeg for audio extraction and conversion to the format required by speech recognition system. After extracting the audio from video files, it writes the file with .wav extension and then provided to speech recognition module for further processing.

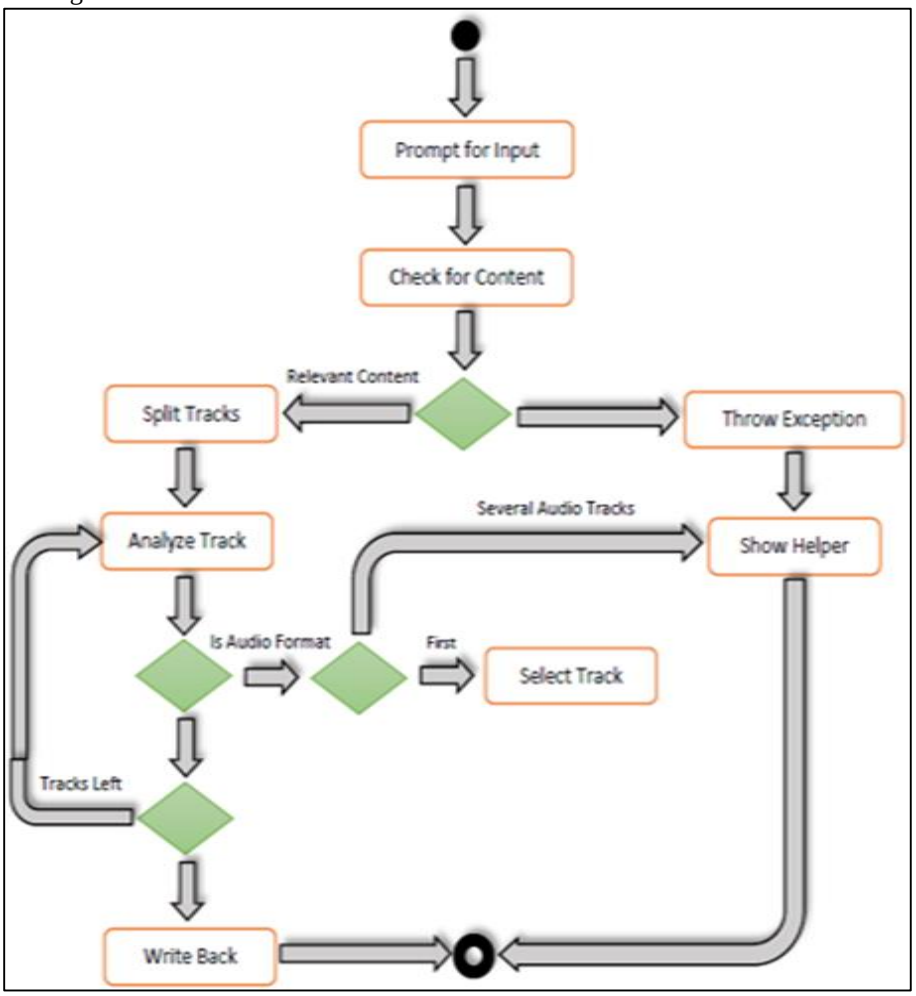

Fig. 2: Audio Extraction

B. Speech Recognition

The .wav file obtained from the audio extraction phase will be passed forward for speech recognition. An open source speech recognition engine called CMU Sphinx will be used in this stage. CMU Sphinx requires following inputs:

- Acoustic Model
- Language Model
- Dictionary file

Hidden Markov Model is used for Speech Recognition for calculating the probability of the occurrence of the words using the acoustic and language model [10]

This system deals with English language videos. Hence, the .wav file generated in previous phase is passed to CMU Sphinx engine along with the US English language model and dictionary file. The text output for the given audio will be generated and forwarded to subtitle generation stage. [2]

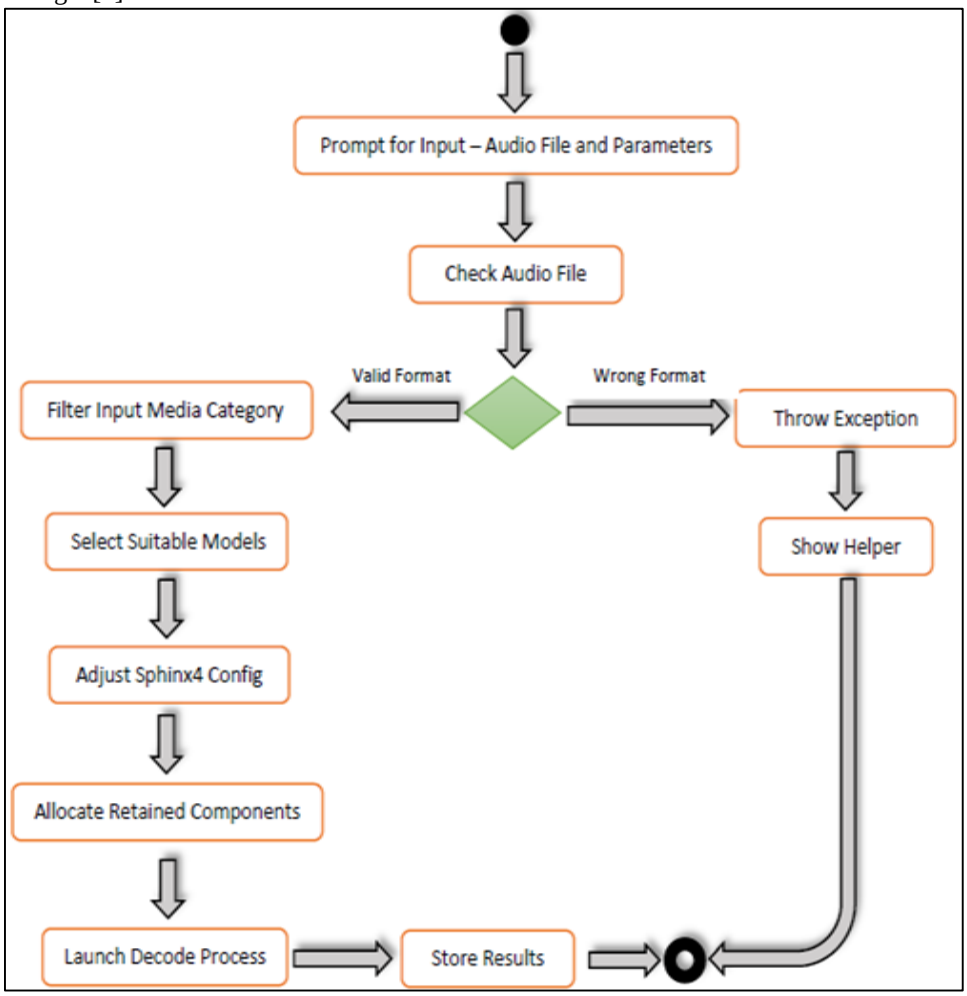

Fig. 3: Speech Recognition

C. Subtitle Generation

This module is expected to get a list of words and their respective speech time-frames from the speech recognition module and then produce a .srt subtitle file. [2] To do so, the module must look at the list of words and use silence (SIL) spoken words as a boundary between for two consecutive sentences. However, we face up to some limiting rule. Indeed, it will not be able to define punctuation in our system since it involves much more speech analysis and deeper design. [1]

```
3
0:00:03,689 --> 0:00:06,969
you an outhouse new a clear comment

4
0:00:06,980 --> 0:00:09,470
american english in only twelve weeks
```

Fig. 4: Sample Subtitle file format

The above figure shows an example of subtitle file generated using the system. Subtitles are text translations of the dialogues in a video displayed in real time during video playback on the bottom of the screen. A typical .srt file has three sections first the Subtitle number indicating which subtitle it is in the sequence. Then the Subtitle timing that the subtitle should appear on the screen, and then disappear and at last the text followed by a blank line to indicate end of current subtitle. [3]

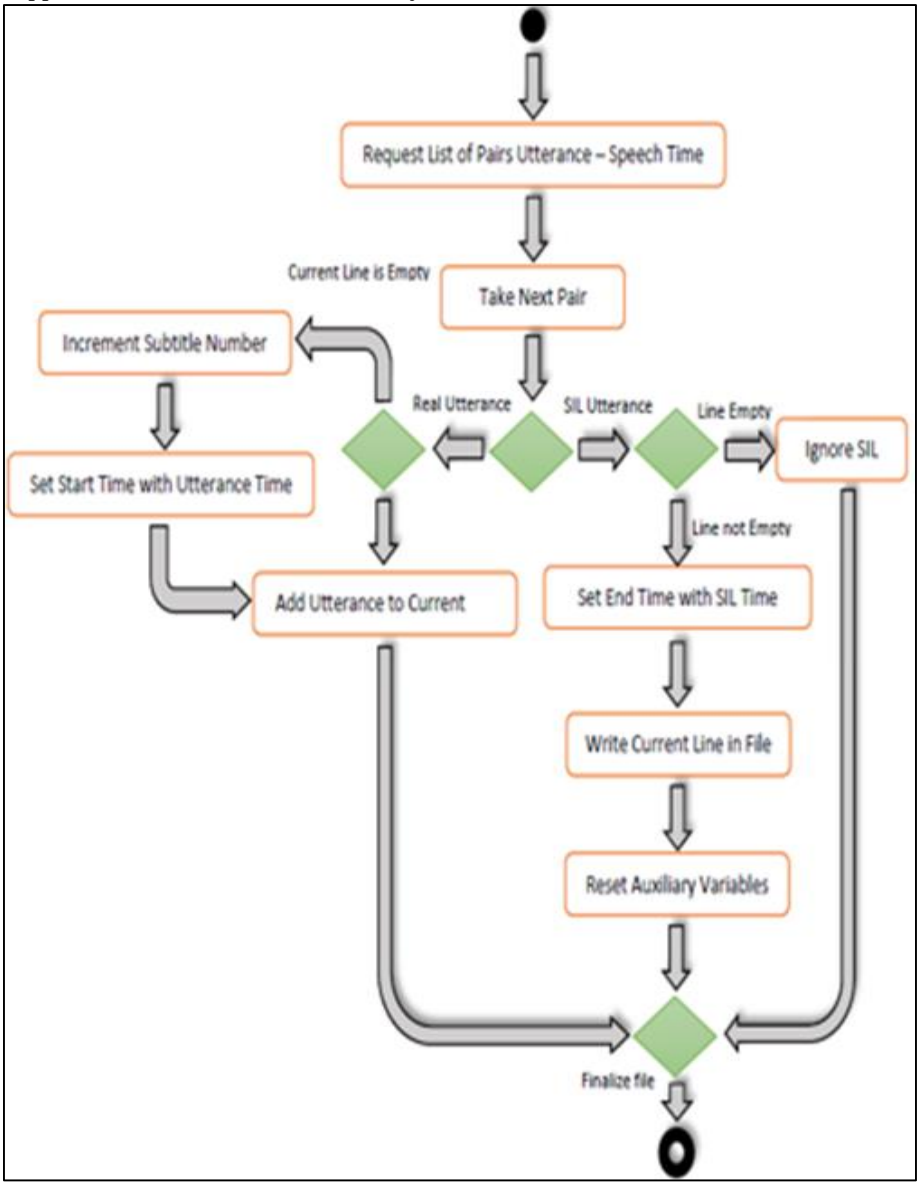

Fig. 5: Subtitle Generation

III. EXPERIMENTATION

The first implementation of the system has been realized on a personal machine of which the characteristics are described in Table

Table 1: Characteristics of the machine used to run the system

OS Name	Microsoft Windows 10
System Manufacturer	ASUS Notebook
System Type	64-bit OS, x64-based Processor
Processor	Intel(R) Core(TM) i3-4010U CPU @1.70GHz
RAM	4.00 GB (3.45 Usable)

Use external software necessary to make the system to run properly. There are mainly three external programs that are needed to run the system and those needs to be systems path or in current directory. These are as follows:

A. FFMPEG

FFMPEG is a free software published under the GNU GPL 2+ license. It is a command-line tool that converts audio or video formats. We are using ffmpeg for audio extraction and conversion to the format required by speech recognition system. For the system to run ffmpeg must be in system path.

B. MPlayer

MPlayer is a free and open media player software. MPlayer can play a wide variety of media formats, namely any format supported by ffmpeg libraries. We used MPlayer into the video player using python libraries that uses MPlayer and it must be in system path to run the system.

C. PocketSphinx

A version of Sphinx that can be used in embedded systems. Pocketsphinx incorporates features such as fixed-point arithmetic and efficient algorithm for GMM computations. We are using it as a Speech recognition system. It must be in system path to generate subtitles automatically and offline.

IV. CONCLUSION

The proposed system is a way to generate subtitles for sound videos. He audio extraction module supplies a suitable audio format that can be used by the intended speech recognition module. Speech recognition produce a list of recognized words and their corresponding time-frames in the audio content. [5] The former list is then used by the intended subtitle generation module to create standard subtitle file that is readable by the most common media players available. The result is the .srt file which contains the text of the input file which is then synchronized with the video. [2]

In a cyber-world where the accessibility remains insufficient, it is essential to give each individual the right to understand any media content. During the last years, the Internet has known a multiplication of websites based on videos of which most are from amateurs and of which transcripts are rarely available. [3] This system work was mostly orientated on video media and suggested a way to produce transcript of audio from video for the ultimate purpose of making content comprehensible by deaf persons.

Thus, the project has been successful according to requirement analysis. Project performs all intended operations correctly and efficiently.

REFERENCES

[1] Boris Guenebaut, “Automatic Subtitle Generation for Sound in Videos”, Tapro 02, University West, pp. 35, 2009.
[2] Abhinav Mathur, Tanya Saxena, “Generating Subtitles Automatically using Audio Extraction and Speech Recognition”, 7th International Conference on Contemporary Computing (IC3), 2015.
[3] Ibrahim Patel1 Dr. Y. Srinivas Rao, “Speech Recognition Using HMM with MFCC- An Analysis using Frequency Spectral Decomposition Technique”, Signal & Image Processing: An International Journal(SIPIJ), Vol.1, No.2, December 2010.
[4] B. H. Juang; L. R. Rabiner, “Hidden Markov Models for Speech Recognition”, Journal of Techno metrics, Vol.33, No. 3. Aug. 1991.
[5] Youhao Yu “Research on Speech Recognition Technology and Its Application,” Electronics and Information Engineering, International Conference on Computer Science and Electronics Engineering, 2012.
[6] Zoubin Ghahramani, “An introduction to hidden Markov models and Bayesian networks”, World Scientific Publishing Co., Inc. River Edge, NJ, USA, 2001.
[7] ImTOO Software Studio, “ImTOO DVD to Video Family”, http://www.imtoo.com/dvd-ripper.html/
[8] Xilisoft Corporation, “Xilisoft DVD to Video Ultimate”, http://www.xilisoft.com/dvd-ripper.html/
[9] Frederick Jelinek, “Statistical Methods for Speech Recognition”. MIT Press, pp. 7, 1999.
[10] Sadaoki Furui, Li Deng, Mark Gales, Hermann Ney, and Keiichi Tokuda,” Fundamental Technologies in Modern Speech Recognition,” Signal Processing, IEEE Signal Processing Society, November 2012.
[11] S. Ross, “Sphinx-4 a speech recognizer written entirely in the java programming language”, pp. 5, 1999-2008. URL http://cmusphinx.sourceforge.net/sphinx4/.