International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 679 - 682

Subtitle Generation: Merging Audio Extraction and Speech Recognition

Anjali Chachra
Department of Computer Engineering
University of Mumbai
anjalichachra24@gmail.com

Anand Singh
Department of Computer Engineering
University of Mumbai
anandsingh1994@gmail.com

Aniket Sankhe
Department of Computer Engineering
University of Mumbai
amsankhe3@gmail.com

Rahul Waghmare
Department of Computer Engineering
University of Mumbai
rahulwaghmare8896@gmail.com

Sonika Nayan
Indian Institute Of Engineering Science And Technology
Department Of Information Technology
sonikanayan@gmail.com

***Abstract*–**In the past few years an outstanding growthhas endorsedthe necessity of videos for the determination of communication. However, non-native diction speakers or people with hearing affliction are unable to take advantage of this robust medium of communication. So subtitles are provided for videos to overcome the problems caused by hearing disabilities or language enclosure. The subtitles generated are stored in the form of a subtitle file most commonly having a .srt extension. There are many softwares available to generate subtitles file manually, however software for automatically generating subtitles are scarce. In this paper, we introduce a system that we have envisioned will generate subtitles automatically through a 3- stage process: Audio extraction, Speech recognition and Subtitle synchronization.Three parts are eminent: The first one includes splitting audio from video and converting the audio in suitable format if necessary. Recognition of speech contained in the audio is second module. Generateing a subtitle file from the recognition results of the previous step is final stage.

Key words: *Audio Extraction, Speech Recognition, Auditory Model, Language Model, Subtitle Generation, Sphinx-4*

*****

I. INTRODUCTION

Video has gotten to be a standout amongst the most prevalent sight and sound ancient pieces utilized on PCs and the Internet. In a larger part of cases inside a video, the sound holds a critical spot. From this announcement, it seems fundamental to make the comprehension of a sound video accessible for individuals with auditable issues and in addition for individuals with crevices in the talked dialect. The most normal route lies in the utilization of subtitles. Notwithstanding, manual subtitle creation is along and exhausting movement and requires the nearness of the client. Wherefore, the investigation of programmed subtitle era has all the earmarks of being a substantial subject of exploration. These days, it exists numerous product managing subtitle era. Some continue on copyright DVDs by separating the first subtitle track and changing over it in an organization perceived by media players, for instance Im-TOO DVD Subtitle Ripper, and Xilisoft DVD Subtitle Ripper. Others permit the client to watch the video and to embed subtitles utilizing the course of events of there video, e.g. Subtitle Editor, and Subtitle Workshop. It can likewise be discovered subtitle editors giving offices to handle subtitle arrangements and straightforwardness changes, for example Jubler, and Gaupol. Along these lines, programming creating subtitles without interference of individual utilizing discourse acknowledgment have not been produced. Along these lines, it appears to be important to begin examinations on this idea.

1.1. Project Description

The current apriorism work principally tends to answer our problematic by presenting a potential system. Three distinct modules have been defined, namely audio ex-traction, speech recognition, and subtitle generation (with time synchronization). The system should take a video file as input and create a subtitle file (sub/srt/sst/son/txt) as output.

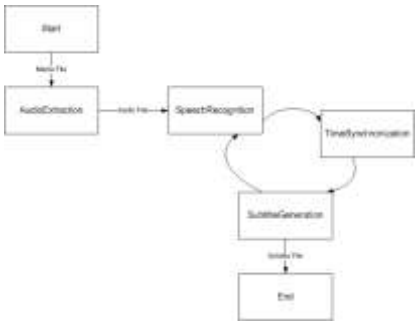
Audio Extraction: The sound extraction timetable is required to give back a reasonable sound arrangement that can be utilized by the discourse acknowledgment module as significant material. It must handle a characterized rundown of video and sound configurations. It needs to confirm the document given in information with the goal that it can assess the expulsion likelihood. The sound track must be returned in the most solid configuration. [6]

Speech Recognition: The discourse acknowledgment calendar is the key part of the framework. Surely, it influences specifically execution and results evaluation. Initially, it must get the sort (film, music, data, home-made, etc...) of the info record as regularly as would be prudent. At that point, if the sort is given, an appropriate preparing strategy is picked. Something else, the normal uses a default arrangement. It must have the capacity to perceive quiets so that content delimitations can be perceived. [3]

Subtitle Generation: The subtitle era routine intends to make and write in a record so as to include different pieces of content relating to word constrained by quiets and their

individual begin and end times. Time synchronization contemplations are of primary significance. [1]

Time Synchronization:A period synchronization strategy as of now exists in the information obtaining framework; we don't have to embed a sync occasion into the information. Synchronized video is urgent for information obtaining and telecom applications. For continuous applications, out-of-sync video may bring about jitter, unevenness and inactivity. For information examination, it is essential to synchronize various video channels and information that are gained from PCM, MIL-STD-1553 and different sources. These days, video codec can be effortlessly gotten to play most sorts of video. In any case, a lot of exertion is still required to build up the synchronization strategies that are utilized as a part of an information securing framework. This paper will depict a few techniques that TTC has received in our framework to enhance the synchronization of various information sources. A standout amongst the most essential information examination errands that must be performed on an information obtaining framework is the relationship of video casings with occasions, sensor information, aeronautics transport information or other video outlines that happen in the meantime. Tragically, the time stamps gave by MPEG-2 Transport Streams is inadequate for synchronizing most information procurement frameworks unless extra data is given. In a few usages, a sync marker is all the while embedded into all information and video channels. TTC's technique works an alternate way; it exploits precise Synchronized time data IRIG time.



II. BACKDROP LEARNING

There were more than a small number of programming languages that could have been used in the formation of AutoSubGen. A quick general idea on the Internet indicated to centre of attention on C/C++ and Java. We now analyse the positive points of the two abovementioned languages based on the study of J.P. Lewis and Ulrich Neumann. On the one hand, C++ provides remarkable characteristics towards speed, cross-systems capabilities, and well-tested packages. On the second hand, Java offers an intuitive syntax, portability on multiple OS and reliable libraries. They both are used in the Sphinx speech recognition engine. Java propose the () API allowing developers to deal with media tracks (audio and video) in a well-organized way. In the article of Lewis, it appears Java performances are relatively similar to C/C++ ones separately of the hardware. Moreover, Java performances should be even higher in theory. The absence of pointers, the use of efficient garbage collector and the run-time compilation play a key role in Java and their enhancement should soon permit to go away from C++ performances.

Java, Sphinx-4 and is the core of the speech recognition part. The subtitle generation uses Java file capabilities.

2.2. Audio Extraction

2.2.1. Fundamentals

A media movie is generally composed of a video track and an audio track. During the collection, the tracks are gathered together and the final artefact is a single file. In this section, we study the feasibility to isolate the audio track from the rest of the file in order to solely process sound. A Google research on the Internet for ‘extract audio from movie’ provides multiple links to software but few tutorial or tip to implement it on his own, especially in Java for this apriorism concern. We oriented the work on the API. This API provides many interesting features for dealing with media objects. The next sub-section describes the different aspects of it. [6]

2.3. Speech Recognition

2.3.1. Fundamentals

A high level overview of how speech recognition works is provided here. The article “A Speech Recognition and apriorism Tool “offers a clear and concise overview of this field while essential methods used in this process are well exposed in “Statistical Methods for Speech Recognition”. This technology permits a computer to handle sound input through either a microphone or an audio file in order to be used to interact with the machine. A speech recognition system can be designed to handle either a unique speaker or a several number of speakers.

2.4Sphinx-4

2.4.1 In a Nutshell

Sphinx-4 is an open source project led by Carnegie Mellon University, Sun Microsystems Inc. and Mitsubishi Electric Research Laboratories. A white paper by Walkeretal presents an overview of the framework. As previously noted, it is

 ISSN: 2321-8169
679 - 682

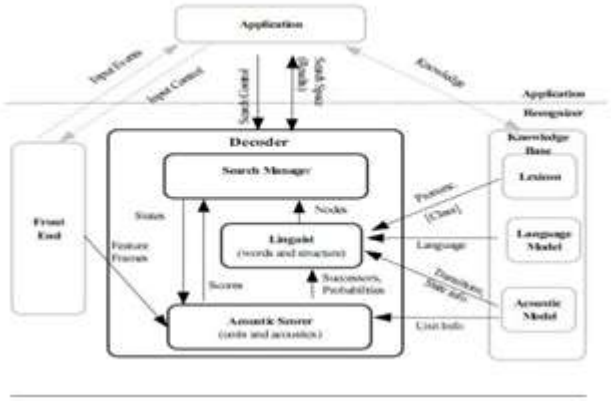
completely written in Java. It offers a highly mixture and flexible architecture as well as adaptable APIs, supports any auditory model structure, handles most types of language models, provides new algorithms in order to get word level hypotheses, and accepts multimodal inputs. [7]

2.4.2 Structural Design

These paragraphs give details the main phases of the Sphinx-4 structural design. We differentiate three principal modules: the Front-end, the Decoder and the Linguist getting material from the Knowledge Base. The Front-end gets a single or several inputs signals and computes them so that a sequence of Features is created. The Linguist generates a Search Graph by translating any kind of standard language model, with the aid of articulation information contained in a Lexicon, also called Dictionary and structural information stored in sets of Auditory Model. The Search Manager component located in the Decoder uses the aforesaid Features as well as the search graph in order to realize the decoding. Results are formed. Controls might be produced to any enabled component and become an effective partner in the process.

III. EXPERIMENTAL SYSTEM

This division represents the experimental system to be setup. First, the requirements as well as the analysis and design of such a system are presented. The second section discusses the implementation phase in which are de-scribed the use of existing software and the creation of customized sections.



3.1. Scrutiny and Aim

3.1.1. Scheme Synopsis

We presented the basic necessities of our system. We now propose a deeper investigation in order to clearly evaluate the real needs. The programming language association in the previous section gave us the desire to mainly use the Java environment to put into practice the system. We also discovered Java components being usable for audio extraction and speech recognition. A smart adaption of those artefacts is essential to offer a suitable system. The subtitle generation will use the different symbols and tags produced by the speech recognizer in order to generate SRT files.

3.1.2. Audio Extraction

This module aims to output an audio file from a media file. It takes as input a file URL and optionally the wished audio format of the output. Next, it checks the file content and creates a list of separated tracks composing the initial media file. An exception is thrown if the file content is irrelevant. Once the list is done, the processor analyses each track, selects the first audio track found and discards the rest (note that an exception is thrown if there is more than one audio track). Finally, the audio track is written in a file applying the default or wished (if mentioned) format. [4]

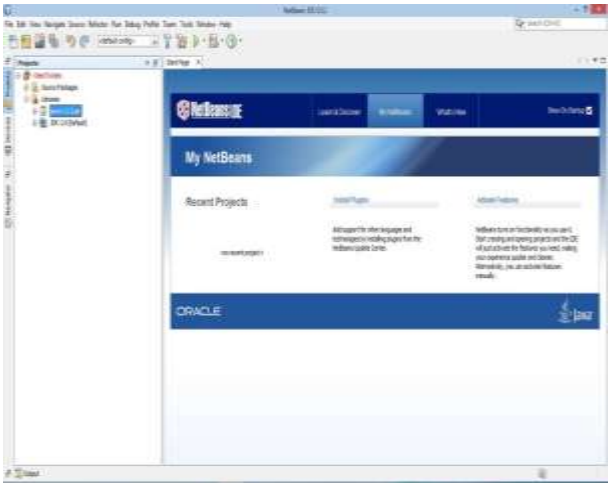
3.1.3. Speech Recognition

According to the limited duration of this apriorism work and because of the complexity of the task, it would not have been feasible to design and realize a new ASR module. In section we talked about the different Sphinx systems. They have been developed for nearly two decades and have proved their reliability within several ASR systems where they were integrated in. It is therefore natural we opted for the Sphinx-4 decoder. As it was introduced in the background study, Sphinx-4 has been entirely written in Java. It was previously talked about the audio formats received by Sphinx. At the instant, it mainly supports WAV or RAW files. It is therefore a complicated task to obtain directly the right format. It often requires various treatments and conversions to reach the abovementioned idea. The component described previously thus provides a way to extract audio from a media file but does not offer conversion tools. Otherwise, it is envisage using the services of VLC. [3]

IV IMPLEMENTAION

The JAVE (Java Audio Video Encoder) library is Java wrapper on the ffmpeg project. Developers can take take advantage of JAVE to transcode audio and video files from a format to another. You can separate and transcode audio and video tracks, you can resize videos, changing their sizes and proportions and so on. Many other formats, containers and operations are supported by JAVE To use JAVE in your Java application, you have to add the file jave-1.0.jar in your application CLASSPATH. JAVE runs on a Java Runtime Environment J2SE v.1.4 the most important JAVE class is it.sauronsoftware.jave.Encoder. Encoder objects expose many methods for multimedia transcoding. In order to use JAVE, you always have to create an Encoder istance:

Encoder encoder = new Encoder();



V. CONCLUSION

We projected a way to generate subtitles for sound videos. An absolute system together with the three required modules introduced in segment could not be realized since the audio conversion needed more resources. VLC gave a correct solution but a custom component coded in Java is expected in further work so that portability and installation of the system is rendered straightforward. However, the expected output for Each phase has been reached. The audio extraction module provides a suitable audio format to be used by the speech recognition module. This one generates a list of recognized words and their equivalent time in the audio although the accuracy is not sure. The former list is used by the subtitle generation module to create standard subtitle file readable by the most common media players available.

In a virtual world where the ease of access remains unsatisfactory, 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 unpaid and of which transcripts are hardly ever available.

ACKNOWLEDGMENTS

We studied this topic as part of our final year project in Computer Engineering. Devotion and strong research work have been formative factors in the writing of this apriorism.

We would like to thank the group of people for giving insightful tips about technical concerns when needed. I also would like to express my thankfulness to-wards my teachers and offering precious advices concerning main issues of such an apriorism work.

REFERENCES

[1] AbhinavMathur, Tanya Saxena, Generating Subtitles Automatically using Audio Extraction and Speech Recognition, 7th International Conference on Contemporary Computing (IC3), 2015.

[2] SadaokiFurui, Li Deng, Mark Gales,Hermann Ney, and Keiichi Tokuda,, Fundamental Technologies in Modern Speech Recognition, Signal Processing, IEEE Signal Processing Society, November 2012.

[3] B. H. Juang; L. R. Rabiner, "Hidden Markov Models for Speech Recognition" Journal of Techno metrics, Vol.33, No. 3. Aug., 1991.

[4] Seymour Shlien,"Guide to MPEG-1 Audio Standard", Broadcast Technology, IEEE Transactions on Broadcasting, December 1994.

[5] Ibrahim Patel1 Dr. Y. SrinivasRao, "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.

[6] Yu Li, LingHua Zhang, "Implementation and Research of Streaming Media System and AV Codec Based on Handheld Devices"12th IEEE International Conference on Communication Technology (ICCT), 2010.

[7] Kai-Fu Lee,IEEE,Hsiao-Wuen Hon and Raj Reddy, "An Overview of the SPHINX Speech Recognition System"IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 38,No.1,January 1990.

[8] http://cmusphinx.sourceforge.net/doc/sphinx4, accessedOctober 2015

[9] http://www.sauronsoftware.it/projects/jave/manual.php