International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

Appellation using Dual Analysis

Beatrice Jeevaraj[1], Rhea Anthony[2], Crystal D'Souza[3], Valencia Joseph[4], Abhishek Jotshi[5]
[1, 2, 3, 4, 5]Xavier Institute of Engineering, Mumbai University, Mumbai, Maharashtra-400016
Email Address: [3]crystal13dsouza@gmail.com

***Abstract*—** Appellation generation using dual analysis is a comparative study of audio and video aspect of a particular file in order to generate subtitles without the use of internet. We can assist people to watch a video and equip themselves with the subtitles without having to do much work.

***Keywords*—** Speech recognition; subtitle generation; audio; video.

I. INTRODUCTION

Visual data has been highly incorporated globally due to the ease of conveying information. It has been used in the field of entertainment, education, personal use, storing data via satellites, surveillance cameras, etc. However due to linguistic barriers, the audio of the video might be illegible. Hence the use of subtitles can help viewers to understand better. Generation of subtitles is a tedious, monotonous, process. It can be done by manually trying to hear and extract the data which might prove to have errors. Also if the video has noise or the data is corrupted, the video might end up being useless. Hence the video should be processed in such a way that the generation the subtitles are more accurate.

II. PROBLEM DEFINITION

As of today, speech recognition of the extraction of text from an audio or video file is yet evolving and getting more advanced. This paper gives information about processing in 2 modules [1]: Processing and Subtitle generation.

III. DESCRIPTION

The working of the system is as follows:

A. Analysis

The analysis is based on two aspects: Audio and Video. The audio which is extracted from the video is given as input for processing to the speech recognition phase. After the speech is converted to text, it is given to the translation module to convert it into another language. This translated text is then given to the subtitle generation module in order to generate the subtitles. For the video aspect, processing it needs the video to be segmented into frames similar to audio extraction. These frames are used for further processing. Based on feature extraction, the processing takes place and gives the output, which can be given for translation, so as to compare both the aspects. This is an attempt to provide an alternative for audio processing.

In the video aspect, region of interest is used to extract the lip and detect the words said. The extracted part is then compared to each frame which is an image in the database. The output of video and audio is then used for a comparative study to find out which one is more efficient.

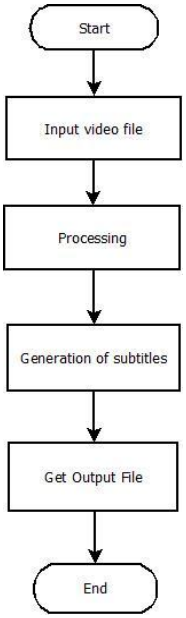

Fig. 1. Flowchart of subtitle generation.

B. Design

The frontend is made in such a way that the actual processing is hidden, maintaining transparency for better aesthetic look, as that the user need not know the entire procedure. Also it is created keeping in mind simplicity, so as to make it easy for the users to utilize. In the backend, we make use of a database to store segmented video and audio that would be used to find comparative values with the existing database in order to find the word that the speaker is saying.

IV. IMPLEMENTATION METHODOLOGY

This project will be divided into two parts:

- Processing
- Subtitle generation

A. Processing

Processing consist of two aspects:Audio and Video. The Audio aspect has two modules in itself: Audio Extraction and Speech Recognition. The Video aspect consists of Feature Extraction [1].

1. Audio aspect

Beatrice Jeevaraj, Rhea Anthony and, Crystal D'Souza, "Appellation using dual analysis," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 235-236, 2016.

a. *Audio Extraction*: The audio needs to be extracted from the video for the purpose of processing [3]. The extracted audio is divided into frames in order to make it easy for the conversion purpose. For simplicity purpose, the spoken words consist only of digits 0 to 9.
b. *Speech to Text conversion:* Speech recognition is the process which is responsible for the conversion of speech to text. The input to this module is the output from the previous module which is Audio extraction. This extracted audio is divided into frames for easy processing. Each frame consists of some values. Every frame will have certain values which will be compared to the values of the frames in the database. Its basic function is to convert the spoken text into a format which can then be processed.

2. *Video Aspect*

a. *Feature Extraction*: For the video aspect, processing it needs the video to be segmented into frames similar to audio extraction. Out of the many frames it is divided into, only the key frames are selected for further processing. Based on feature extraction, the processing takes place and gives the output. In the feature extraction phase, the lip of the person is extracted from the frame, in other words detecting the region of interest (ROI) [2], and compares with the database on the basis of similarity values to the frame images, of which lip portion is extracted. This output is given for translation, so as to compare both the aspects. This is an attempt to provide an alternative for audio processing. The output of video and audio is then compared to find out which one is more efficient.

B. *Subtitle Generation*

The output generated from the above module, that is, Speech Recognition, contains the words spoken in the audio file, will be saved in a text file. This file can be translated using an API. The translated output is then updated into another file which is converted to a .srt file. This will be the subtitles which is the final output.

V. CONCLUSION

This paper, implementing the methods mentioned can be of help to people who have auditory challenges, to people who have difficulty due to linguistic barriers and also when the audio might be corrupted.

REFERENCES

[1] A. Mathur, T. Saxena, and R. Krishnamurthy, "Generating subtitles automatically using audio extraction and speech recognition," *IEEE International Conference on Computational Intelligence & Communication Technology*, pp. 621-626, 2015.
[2] P. Sujatha and M. Radhakrishnan, "Speaker –Independent visual lip activity detection for human-computer interaction," *IJRET: International Journal of Research in Engineering and Technology*, vol. 2, issue 11, pp. 561- 562, 2013.
[3] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE Transactions on Cybernetics*, vol. 44, issue 2, pp. 2, 2014.

Beatrice Jeevaraj, Rhea Anthony and, Crystal D'Souza, "Appellation using dual analysis," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 235-236, 2016.