IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Transcripter-Generation of the transcript from audio to text using Deep Learning

Fatima Ansari[1], Ramsakal Gupta[2*],Uday Singh[3], Fahimur Shaikh[4]

Department Of Computer Engineering, M.H.Saboo Siddik College of Engineering, Mumbai

Corresponding Author: ramsakal7055@gmail.com Mob: +91 8689968727

Available online at: www.ijcseonline.org

Accepted: 26/Jan/2019, Published:31/Jan/2019

Abstract— A video is the most powerful medium in the propagation of information and important part of the video for exchanging the information is audio, which is an important aspect of the video on which the whole message depends and as it is used in all field like Teaching, Entertainment, Conference Meeting, News Broadcast. So converting the Audio into Text in Documented format make easy for referring purpose as it is difficult to search the said word in the video as compared to the transcript. The main objective of developing this system is to present an automated way to generate the transcript for audio and video. As it is not possible to make the same informative video in all Languages. So this the place where our System plays an important role. It will extract the audio from the given video and transcript is generated based on which it can be translated into any desired language. It can be very useful for people who speak the language which is not used by the majority of the population. In this way, it has much application in all field where information exchange is happening based on Video

Keywords: Neural Network, Audio extraction, Speech recognition, Time synchronization, Automatic Transcript generation, Natural language processing, Connectionist Temporal Classification (CTC), Hidden Markov Model (HMM).

I. INTRODUCTION

Multimedia content such as video and audio files may be supplemented with closed captions for accessibility. Captioning multimedia content is typically a two-step process:

1) transcript the content to obtain text.

2) Realign the transcription with the content to produce closed transcripts text.

The closed transcripts text are the text translation appearing on the screen of a video displayed in real time during its playback. captions and transcripts can make multimedia content accessible and also improve the "searchability" of the content. The users of the transcript are the people unfamiliar with the language, deaf and those trying to improve the reading skills of the language. Depending as per the need, the language of the transcript could be the same as that of the video, or another one suitable to the user of reliable speech conversion for spoken document/information retrieval is a challenging problem when data is recorded across different media, equipment, and time periods and at the same time it contains unwanted noise and in the case of songs, music has to be separated from lyrics mean audio classification is required.

In this paper, we present a system that takes a raw transcription of a video as input and generates: (i) accurately time-aligned closed captions; and (ii) a readable, well-formatted transcript with punctuation, capitalization, and paragraph segmentation. Because the manual effort is reduced to straight transcription, considerable time and money can be saved and more multimedia can be made accessible and searchable. lecture. In this Local preference of the keywords are reflected with a cache model by referring to slide used during each utterance.

II. LITERATURE REVIEWS

We have gone through the various system that works in the speech recognition section. we reviewed many available papers in the market and evaluated them efficiently. In paper [1]

Authors have worked on classification of the video using the speech recognition. They have used the Convolutional Neural Network(CNN) for their system.

The Author in [2] is proposed a language model adaptation for their system to transcript the video lecture based on the presentation slide that was used in the

The aim of the author in [10] is to develop a system that filters the voice from the mixed audio of a conference. In this, the author has used the Recurrent Neural Network (RNN) model for the transcription of the audio. They have used CTC(connectionist Temporal Classification) algorithm for mapping a sequence of audio to a sequence of character.

We studied other papers which were more traditional in approaches and used HMM model. In that papers, they proposed a system for punctuation also[12].

In paper[5-9] they are more on some specific topic recognition that is like educational video classification, conference transcription using slides.

In paper[3-4] The author insists more on working on a translation of the recognized language.

We want to develop a system that transcript the audio to text with accuracy that is enough to make it a practical solution in real life.

III. IMPLEMENTATION OF OUR SYSTEM

Flowchart Diagram of the working of the system is as shown in the Fig.1. As shown first we will get the input in terms of the audio/video file. If the file is in video format then we have to first extract the audio from the video. Once we get the audio then we have to divide it into the chunks to feed the model and list of the probable word is generated and the most probable word is selected the selection will be based on the past input as it happens in RNN model. We have to repeat this process until all the chunks are over.

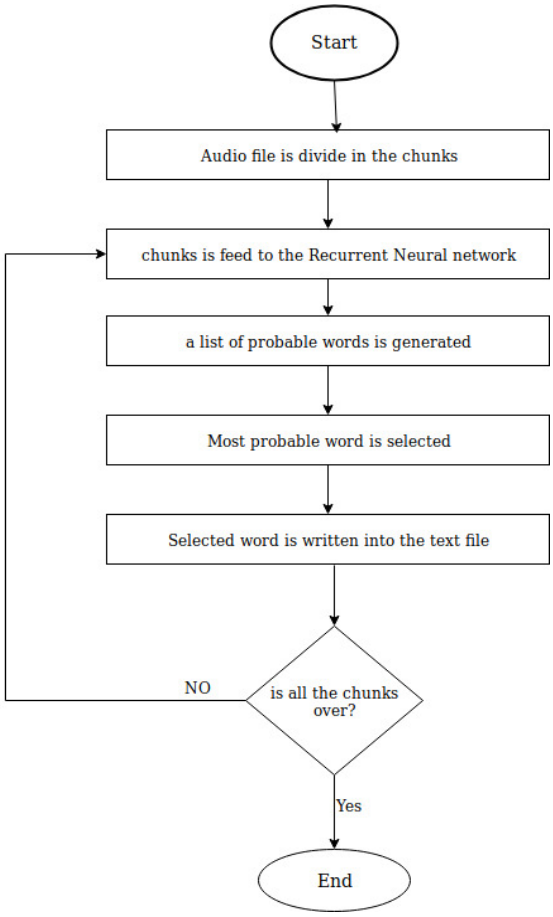

Figure 1. Flowchart of the proposed system.

A. Connectionist Temporal Classification.

In Connectionist Temporal Classification (CTC), a sequence of audio is mapped to a sequence of characters. In this, there is no knowledge about how the length of the input is aligned along with the output length. This is more favored as it is not easy to specify how the alignment should be. For example, specifying a certain window of the sound wave to align with a character will not work for speakers with different rates of speech CTC circumvents the alignment problem by simply assigning a token to every input step and collapsing the string of tokens. For a value X, it gives an output distribution for all possible Y's, which will be used during inference to predict a most likable output. To train the model, we maximize the conditional likelihood of the data with gradient descent. Forming the intermediate alignment is done by assigning a token to each input step. The tokens can be any character a,b,…,z blank. It will assign a null or blank token account where no word is said for periods of time and it is silence and hence no corresponding output. To collapse the alignment, we remove any repeated characters not separated by a blank token. This allows for repeated characters. For example, the algorithm might assign the sequence "CCEE-LL-L-OO" to the input, where the "-" character represents the blank or null token. This sequence is then collapsed to "CELLO". The characters are chosen based on a distribution given by the network so using this per time-step output we can compute the probabilities of different alignment sequence[3]

B. System Architecture

Component Diagram of the system is shown in Figure 2.This system consists of three modules which are audio extractor, RNN model and output of the system.

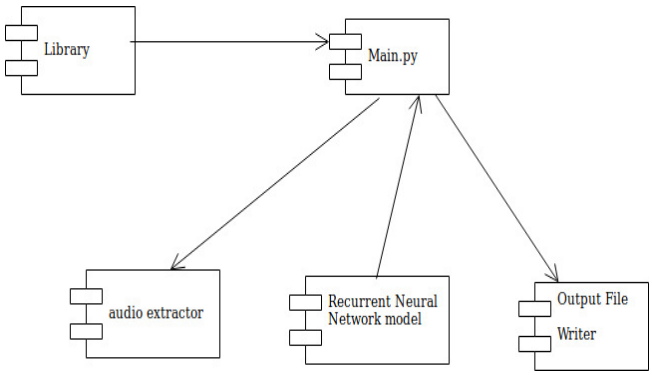

Figure 2 Component Diagram

To understand the basic working of our system Use case diagram is shown in figure no.3.This show requirements of a system including internal and external influences. These requirements are mostly design requirements.

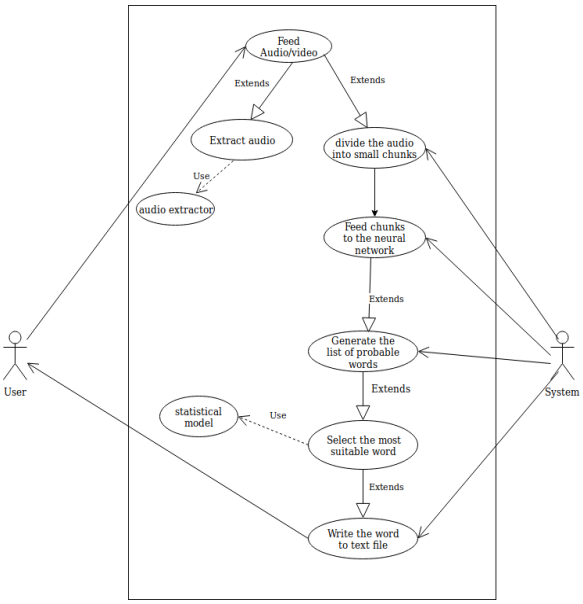

Figure 3 Use case diagram for the proposed system

IV. CONCLUSION

By using Transcripter, a file will be generated for any English videos and audios. This system will minimize the efforts for manually transcripting the file which is a very tedious task. Anyone will be able to generate the transcript file as this system is very easy to use and just needs input which can be provided by anyone through a graphical interface. Also, this system can be used with an online website to provide videos along with the transcript. Many types of formats of video and audio will be supported by this system.

Thus, we propose a novel system for taking input in the form of recorded audio and video (voice) signal and translating the voice signal to provide an output in different multiple Indian languages. This system provides an option of translating the recorded voice signal in languages like English, Hindi, Tamil and Telgu which are among the most spoken languages in India. This may lead to great help to the people lacking the power of understanding the particular Indian language and it will be helpfull to communicate with the people belonging to the different region of the country.

V. FUTURE SCOPE

While speech recognition is performing very good for most applications today such as personal digital assistants like Google, Siri, Alexa. Many works still have to be done to turn the automated transcript generation into a working system. This is because the results are being examined by a human as compared to a machine which is time-consuming. A full transcript generated from a 20-minute conference might give errors while using the best speech recognition engines. So, this software has a huge scope for improvement and extension in terms of optimization and applications. It can be made to translate into multiple Indian languages using the same input structures. It can be made much more efficient and optimize using various efficiency enhancing and filtering algorithms like the Wiener Filtering and Signal subspace approach (SSA) model and many others. This system can also be used as hearing aid to enable the same service for people who are deaf as well. This technology is different from others as it enables speakers of different languages to communicate with each other and adds value to humankind in terms of World Peace, Science and Commerce, Cross-Cultural exchange, World Politics, and Global Business. In future work, we could add speaker change detection, acoustic event detection and music identification to the preprocessing and alignment stage of our system

REFERENCES

[1]Houssem chamber, Marlon Oliveira, Kevin McGuinness, Suzanne Little, Keisuke Kameyama "Educational video classification by using a transcript to image transform and supervised learning."

[2] Tatsuya Kawahara, Yusuke Nemoto, Yuka Akita. "Automatic lecture Transcription by exploiting presentation slide information for Language Model Adaption."May 2008 IEEE

[3] Wai Fong Chua, "Teaching and learning only the language of numbers—monolingualism in a multilingual world, " Critical Perspectives on Accounting, Vol. 7, No. 1, pp. 129-156 February 1996, ISSN 1045-2354, http://dx.doi.org/10.1006/cpac.1996.0019.

[4] G. Nowak, S. Grabowski, C. Draus, D. Zarebski and W. Bieniecki, "Designing a computer-assisted translation system for multi-lingual catalog and advertising brochure translations," in Proc. of Sixth International Conference on Perspective Technologies and Methods in MEMS Design pp.175- 180, 20-23 April 2010.

[5] Houssem chamber, Marlon Oliveira, Kevin McGuinness, Suzanne Little, Keisuke Kameyama, Paul Kwan, Alistair Sutherland, "Educational video classification by using a transcript to image transform and supervised learning." 13-15 July 2016

[6] J. Santos and J. Nombela, "Text-to-speech conversion in Spanish a complete rule-based synthesis system," in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.7, pp.1593-1596, May 1982.

[7] F. Y. Sadeque, S. Yasar and M. M. Islam, "Bangla text to speech conversion: A syllabic unit selection approach," in Proc. of International Conference on Informatics, Electronics & Vision (ICIEV), pp.1-6, 17-18 May 2013.

[8] Xia Linsi, N. Yamashita and Toru Ishida, "Analysis on Multilingual Discussion for Wikipedia Translation," in Proc. of Second International Conference on Culture and Computing, pp.104-109, 20-22 Oct. 2011.

[9] Tatsuya Kawahara, Yusuke Nemoto, Yuka Akita, "Automatic lecture Transcription by exploiting presentation slide information for Language Model Adaption." August 2008.

[10] Nikolas Lee, Jia Wern Yong,"Automated Transcript Generation for the video conferences" 11-Nov 2017

[11] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in Proceedings of the International Workshop on Spoken Language Translation, 2011.

[12] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 474–485, 2012.

Authors Profile

Ansari Fatima Anees Works as an Assistant Professor at M.H. Saboo Siddik College of Engineering in Computer Engineering Department.
Qualifications: - B. E. (IT) M.E. (IT), Also, having teaching experience of more than 10 years in this institute.

Ramsakal Gupta. Currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Computer Engineering Department.

Uday Singh. Currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Computer Engineering Department.

Fahimur Shaikh Currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Computer Engineering Department.

