Movie Captioning For Differently Abled People

Prasad Parsodkar[1], Pradnya Shinde[2], Sangeeta Kurundkar[3]

[1]Prasad Parsodkar, B.Tech, Vishwakarma Intitute of Technology, Pune
[2]Pradnya Shinde, B.Tech, Vishwakarma Intitute of Technology, Pune
[3] Dr.Sangeeta Kurundkar, Dept. of Electronics Engineering, Vishwakarma Intitute of Technology, Pune, Maharashtra, India

---***---

Abstract - *An Entertainment is the most significant part of human considering the aspiration of relaxation or leisure. Regional, cultural and traditional dimensions decide the types of entertainment one would like to have. Drama, music, flow of ideas, and in the broader aspect current status of society has brought and wrapped into very unique category of entertainment known as Movie. Movie is the popular and effective form of entertainment and information. It is very disheartening to know that some people in our surrounding has not easy access to Movie as being easy approachable This paper provides the smart helping hand to visually and hearing impaired people who are deprived from the easy interpretation of movie. Concerned paper proposes the flow of processes to caption movie. Flow consist of Automatic Speech Recognition(ASR), Object identification, image and video annotation and parallel working on the audio part of movie using various frequency separation algorithms. This paper hopes that they can watch movie independently without any further assistance.*

Key Word: Automatic speech recognition, object identification, audio separation

1. INTRODUCTION

Since many years entertainment [1] has been very famous form of activity which holds the attention of people whether it may be in the form of dance, dramas, sport, news, storytelling and different performances included in culture. The regional or local art-form conspires of concerts, stage magic, festivals devoted to dance and many more. The process of being entertained is come under the association with the amusement [1]. The definition of amusement has been changing in every decade as the trends are taking steps to different directions of ideas.

Movie is being the best category of the entertainment where each and every part performances can be gathered. Movie symbolizes the combination of drama, dance, music, and the vision of director whose spectacles are being used by the audience. Movie is emerge as the entertainment in recent 2 decades but has affected the audience than any other entertainment type can do. Depictions of movies are pretty clear and the ideas to be conveyed are poured into dialogues of the characters of the drama in movie. As we are moving forward movies are seen not only as the entertainment but also the means of flow of ideas in to the society. Today movie is considered as the reflection of society. It gives exact picture of situation of various happening in the surrounding. Many modern notions, methodologies, trends, traditions are benchmarked by the movies. Need of the change in society is undisputedly satisfied by movies. It is the powerful tool for culture, education and leisure. In the list of literature, movie has been added as it has come up as the bucket of things associated with human beliefs.

We can never imagine that to get entertained requires an aid. But it happened to many people if we properly search for them. It is very disheartening to hear that. So to tackle this issue we come up with the method of Movie captioning for those who are deprived of the access to movie. Hearing and visually impaired will be the beneficiaries of this system. Hearing impaired will have the aid of watching, whereas visually impaired will have the aid of audio assistance.

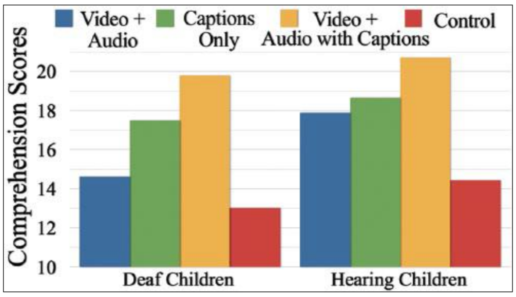
Captions [2] will also benefit everyone who watches videos, from younger children to older adults. They are particularly for the persons watching videos in their non-native language, and persons who are hearing and visually impaired. They are generated via automatic speech recognition, when auto-generated caption reach parity with human-transcribed caption, technology will be able to harness the power of caption.

Our system will consist of Automatic speech recognition, object identification, these are used for the automatic generation of subtitles and to describe the scene as the objects are identified in any particular scene. Next steps led to cocktail party algorithm which will only concentrate on specific frequency. Further this goes into emotion detection from the facial expression and from speech. Mentioned technologies are kept in the black box. The input movie is processed through this black box and taken as the captioned movie. We hope this system will help different abled people to enjoy the movie irrespective of what inabilities they have.

2. Related Work

The early 20th century's golden age of cinema had created alevel playing field for D/deaf and hard of hearing viewers. Silent films, with their interwoven screens of captions (called intertitles), created "the one brief time that deaf and hard of hearing citizens had comparatively equal access to motion pictures"(Schuchman, 2004, p. 231). But in the late 1920s, as talkies (films with synchronized speech) pushed out silent films, the D/deaf community was shut out. Captions began appearing on television shows in the 1970s (with their earliest appearances on ABC's Mod Squad and PBS's The French Chef; Withrow, 1994). In the 1980s, a handful of television shows began displaying captions in real time [3](e.g., the launch of the space shuttle Columbia and the acceptance speeches at the Academy Awards; Block & Okrand, 1983). By the 1990s, captions on TV shows were mandated bythe U.S. law (Erath & Larkin, 2004) [3]. The Twenty-First Century Communications and Video Accessibility Act of 2010 require that captioned TV shows also be captioned when displayed on the Internet.

Chart -1: This diagram explains the statistical data representation using bar graph so that the need of our system is symbolized and the audience of our system will greatly benefited and help them to relax their senses and enjoy the movie.



Sony has personalized the closed captioning concept with the invention of Entertainment Access Glasses. Instead of reading captioning on the screen, these glasses project captions in the air in front of the viewer. The text can be seen by the user, glasses are large enough to fit over conventional eyeglasses.

3. Proposed Algorithm

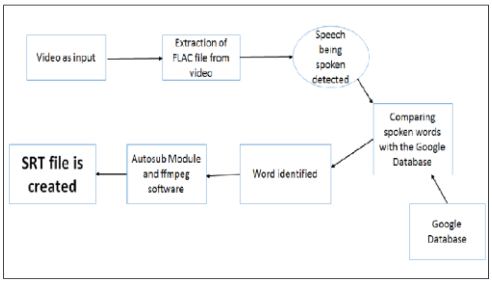
Our proposed algorithm has some flow which will go in the following way where the Automatic speech recognition is the 1st step.

3.1 Automatic Speech Recognition (ASR):

Speech recognition is also known as automatic speech recognition or computer speech recognition which means understanding voice of the computer and performing any required task or the ability to match a voice against a provided or acquired vocabulary. ASR can be also used in bidirectional way i.e. Speech-to-text (STT) and Text-to-Speech (TTS) [4, 5, 6].

Speech recognition module of python gives the platform to convert spoken words into speech. This module is backed by Google speech API. It has its own pre-trained dataset and on that basis the conversion is taken place [14].

Flowchart -1: In this figure the flow of automatic subtitle generation is depicted in which the out is SRT file which is Speech Recognized Text .It has used the Autosub module of python. It keeps the Google dataset as the reference and the uttered word are gathered into the text file.

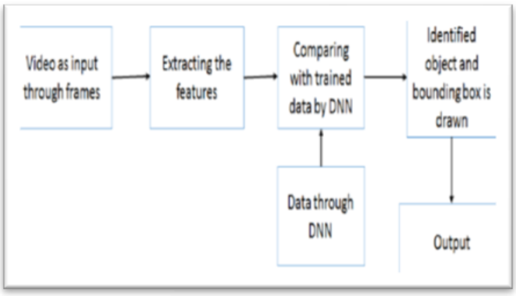

Video as input
Extraction of FLAC file from video
Speech being spoken detected
Comparing spoken words with the Google Database
Google Database
Word identified
Autosub Module and ffmpeg software
SRT file is created

Autosub [6] is a tool for automatic speech recognition and automatic subtitle generation. Video is taken as input from the FLAC file extracted which is Free Lossless Audio Codec file [13]. It has very minute details of audio signal that's why it can't be played how we play .mp3 file. This file consist of MFCC [11] features in the audio, those are compared Google pre-trained dataset and the actual word is estimated. Table-1 has explained it in very jest of the glance[7].

3.2 Object Identification:

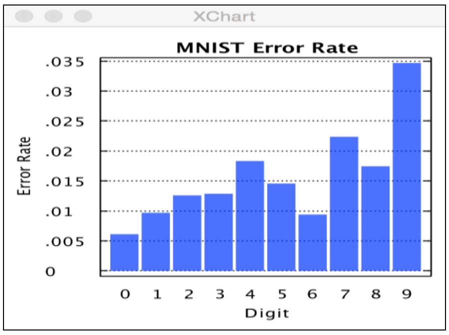
Object identification is very crucial step in our proposed algorithm which is totally based on the machine learning concept. If we provide the labeled data to the Deep Neural Network, it adjusts the weights and bias. Hence the data set is trained so to test it on tested data. We have used 15 classes dataset and can identify 15 objects [8] in all. Along with that object identified are bounded and the percentage of similarity of show at the right top corner of the bounding box. This gives very clear picture of what happenings are taking place in the scene [15].

Flowchart -2:


Video as input through frames
Extracting the features
Comparing with trained data by DNN
Identified object and bounding box is drawn
Data through DNN
Output

The dataset which is trained surely will have the errors as no dataset is absolutely correct as the labeling of the dataset matters a lot on which the DNN [9] is being trained. While going through the dataset of different objects, labeling can have different opinions so as to tackle this problem we can have two-three labeling experts. Some error rate are show in chart 2. It shows the digit-wise error rate and represented in the graphical way. It is concluded that digit 9 has maximum and digit 1 has the least error rate i.e. 0.035 and 0.006(approx.) [10]

Chart -2: Error rate is shown graphically where digit 9 has maximum and digit 1 has lowest.



3.3 Processing on wav file:

Cocktail party algorithm is the speech separation algorithm and mostly used to track the audio frequency in the sound to be heard. Consider there is party going on, having mixture of different sound such as speech, music, background noise and silence. Now our task is the separate the speech from the whole audio file. Here we first get the actual frequency of the speech to be tracked and in every audio chunk that component is detected and the weight of speech in context of the whole music file is maximized so that it gets mire emphasize.

This algorithm has two models in it.

(1) Mixing Model

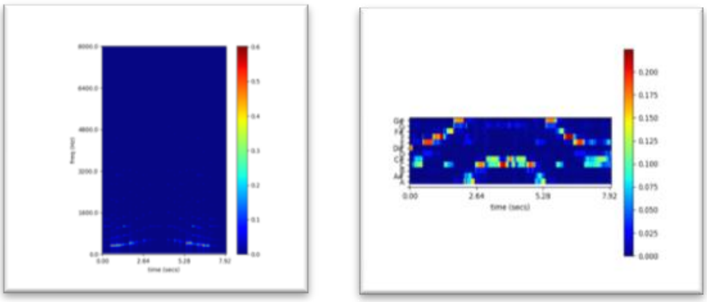
(2) Separation algorithm

3.4 Emotions From Speech:

It is the process of extraction of features from audio file followed by the training and then the testing on unlabeled data. This process uses the PyAudioAnalysis module of python for extraction of MFCC [11] components from that training data. MFCC are Mel-Frequency cepstral coefficient which form a cepstral representation where the frequency bands are not linear but distributed according to the Mel-scale. Short-term and mid-term analysis is has the aforementioned list of features to be extracted the audio signal is first divided into short-term windows (frames) and for each frame all 34 features are calculated. This results in a sequence of short-term feature vectors of 34 elements each.

Another common technique in audio analysis is the processing of the feature sequence on a mid-term basis, according to which the audio signal is first divided into mid-term windows (segments)[11], which can be either overlapping or non-overlapping[12]. Tempo-related features consist of automatic beat induction i.e. the task of determining the rate of musical beats in time is a rather important task, especially for the case of music information retrieval applications.

Chart -3: These graphs show the result of plotting of Time Vs. frequency graph in which we are only concern about the human frequency component denoted by light shades.



It performs four types of functions which are mentioned below:

- Classification: Supervised knowledge [11] (i.e. annotated recordings) is used to train classifiers. A cross-validation procedure is also implemented in order to estimate the optimal classifier parameter (e.g. the cost parameter in Support Vector Machines or the number of nearest neighbors used in the k-NN classifier [11]). The output of this functionality is a classifier model which can be stored in a file. In addition, wrappers that classify an unknown audio file (or a set of audio files) are also provided in that context.

- Regression: models that map audio features to real-valued variables can also be trained in a supervised context. Again, cross validation is implemented to estimate the best parameters of the regression models.

- Segmentation: The following supervised or unsupervised segmentation tasks are implemented in the library: fix-sized segmentation and classification, silence removal, speaker diarization and audio thumb nailing. When required, trained models are used to classify audio segments to predefined classes, or to estimate one or more learned variables (regression) [11].

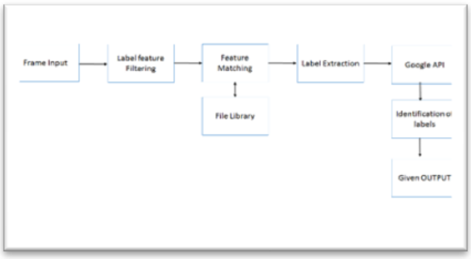
- Visualization: Given a collection of audio recordings PyAudioanalysis can be used to extract visualizations of content relationship between these recordings.

3.5 Image and Video Annotation:

Till now we get many significant information from audio part of the movie. But the same, rather more amount of information can be collected from the video content of the movie. Annotation of image or any video plays very important role knowing the happenings in the scene.

Annotations are the one word description of the image provided, sufficient to get the rest of the information from every frame. In the same way a video can also be annotated there are three level of annotation of video as per the during of video content (1) Frame level: This is nothing but the labeling of single image of the video if it is converted into number of frames.(2)Shot level: Every shot taken the cinematographer has some meaning in itself so those are annotated very easily. Moreover, the meaning of the shot would be the label of the video. (3)Video level: This will take whole video as input but any video whose duration is more than a shot so how a single label can be given to whole video. Still it has some results to show correctly.

Flowchart -2:



3. CONCLUSION

The beneficiaries of the proposed system are mostly for hearing and visually impaired people. As the system built on best suitable technologies and algorithm, it become very user-friendly to access movies. The proposed algorithm generate image captions for visually impaired and audio assistance for hearing impaired. Our results have implications towards how to better adapt existing captioning system.

REFERENCES

[1] C. Barathi and C.D. Balaji,"Trends And Potential Of The Indian Entertainment Industry- An Indepth Analysis", Journal of Arts, Science and Commerce E-ISSN 2229-4686

[2] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hastyand Steven Landau. 2015. Guiding Novice Web Workers in Making Image Descriptions Using Templates. ACM Transactions on Accessible Computing 7, 4: 1–21. https://doi.org/10.1145/2764916

[3] Morton Ann Gernsbacher,"Video Captions Benefit Everyone", Policy Insights from the Behavioral and Brain Sciences2015, Vol. 2(1) 195–202.

[4] M.A.Anusuya and S.K.Katti, "Speech Recognition byMachine: A Review", (IJCSIS) International Journal ofComputer Science and Information Security, vol. 6, no. 3, pp.

[5] Preeti Saini and Parneet Kaur, "Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology –Volume4Issue 2- 2013

[6] AmirsinaTorfi,"SpeechPy- A library for Speech Processing and Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(1): 52–59, 1986.

[7] Joseph P Campbell. Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9):1437–1462, 1997

[8] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S.Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, "Towards fully autonomous driving: Systems and algorithms," IEEE Intell. Veh.Symp. Proc., pp. 163–168, 2011.

[9] Huieun Kim, Youngwan Lee, ByeounghakYim, Eunsoo Park, Hakil Kim," On-road object detection using Deep Neural Network",2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)

[10] Dan Cires¸an, Ueli Meier and J¨urgenSchmidhuber, " Multi-column Deep Neural Networks for Image Classification" , volume 2766 of Lecture Notes in Computer Science. Springer, 2003. 1, 2

[11] PatriarchouGrigoriou and Neapoleos," pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis", PLoS ONE 10(12): e0144610. doi:10.1371/ journal.pone.0144610

[12] Emmanuel Vincent, RémiGribonval, and CédricFévotte," Performance Measurement in Blind Audio Source Separation", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING

[13] D. Impiombato, S. Giarrusso, T. Mineo, O. Catalano, C. Gargano, G. La Rosa, F. Russo, G. Sottile, S. Billotta, G. Bonanno, S. Garozzo, A. Grillo, D. Marano, and G. Romeo, "You Only Look Once: Unified, Real-Time Object Detection," Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip., vol. 794, pp. 185–192,2015.

[14] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. Recent advances in deep learning for speech research at microsoft. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8604–8608. IEEE, 2013.

[15] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on1473–1482.