Convolution Neural Network Implementation in Optical Speech Recognition

G Annapoorni[1], Sayantan Chatterjee[2], Kaustubh Narkhede[3], Gokul Rajes P[4], Saksham Alag[5], N Sricharan phanindra[6]

[1]*Assistant Professor, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[2]*Undergraduate student, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[3] *Undergraduate student, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[4]*Undergraduate student, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[5] *Undergraduate student, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[6] *Undergraduate student, Department of Electronics and Communication, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*

ABSTRACT

Our main motive in this section is to eliminate the challenges which we faced in audio speech recognition and to upgrade the security of audio speech recognition system. so we propose the idea of optical speech recognition using convolution neural network which is used for extraction of visual features. In order to achieve this purpose, we have utilized the concept of CNN which can be explained as trainable modules to extract the required features. The CNN play an important role by processing the input image and segregate the various modalities from the visual image and this extract tells us about the challenges of dependable talk insistence framework from optical data just, without utilizing sound pennant. Our procedure consolidates a camcorder and a ultrasound envisioning framework for meanwhile perceiving the speaker's lips and the adaptability of our tongue. Strikingly, the assertion capacity is few percent lower than which was picked up unwinding sound pennant which improves it for accommodating optical talk certification structure.

Keyword: - *CNN-convolutional neural network, HMM- Hidden Markov model, GMM- Gaussian Markov Model, and Architecture etc….*

---

1. INTRODUCTION

Speech is exceptional among the highest imperative medium by which a correspondence can happen. With the creation and broad utilization of mobiles, phones, information stockpiling gadgets and so on has given a noteworthy help in setting up of discourse correspondence and its breaking down. The term and the essential idea of discourse recognizable proof was started in the right on time with investigation into voiceprint examination which was to some degree like unique mark idea further development and headway in the field of discourse acknowledgment, the people who are physically tested, for example, visually impaired and hard of hearing can undoubtedly speak with the machines. So in natural terms a discourse that is being produced through trachea will be decoded by mind. Lip

inspecting is used to understand or translate talk with no use of listening it, a methodology especially aced by people with listening annoys. The ability to lip interpret engages one by a party insufficiency to chat with other people and to partake in various activities, which by and large would be troublesome. A late advance in the fields of PC vision, plan declaration, and standard overseeing has induced a creation vitality for robotizing this testing undertaking of lip examining.

Unquestionably, modernizing the human ability to lip read, a system proposed as visual talk request (VSR) (or now and again talk exploring), could show the section for other related applications. VSR has gotten a ton of thought in the latest decade for its potential use in applications, for instance, human-PC connection, broad media talk affirmation (AVSR), speaker request, talking heads, development based correspondence declaration and video perception. Its standard point is to see spoken word(s) by using only the visual flag that is made in the midst of talk. Thusly, VSR deals with the visual space of talk and merges picture planning, man-made thinking, object locale, plan authentication, exact outlining, etc. There are two explicit essential ways to deal with oversee accord by the VSR issue, the approach and the complete structure, each having own characteristics and insufficiencies.

Generally, examinations based on optical speech recognition (all things considered from lips flags basically), separate this issue in dual unique stages: the analysis of visual features from grungy pictures, and the classification itself. For the fragment analysis organize, distinctive frameworks have been proposed among which dynamic shape show [3], dynamic appearance show [4], DCT change [5], and fundamental parts examination (PCA) [6]. Besides to sound Based talk attestation, the segmentation plan is an extraordinary piece of the time tended to using models organized to consider the dialog fragments. Among various approaches, graphical models, for instance, covered Markov show up (Gee) [7], Coupled-Gee [8], and dynamic framework [9] has been everything considered inspected. In our past work ([1, 10]), we have searched for after a comparable framework, by using a PCA-based decay for encrypting each the visual, USI pictures, and a dual-types Well Gaussian mixture model classifier. Profound neural systems have appeared numerous areas their capacity to gain portrayals straightforwardly from the crude information and can be utilized to extricate a lot of discriminative highlights. With regards to picture handling, one amazing profound design is the supposed Convolution Neural System (CNN) [11]. The utilization of CNN for signals acknowledgment visually which has been suggested in couple of ongoing examinations.

Here, we acknowledge using CNN to bring out optical includes straightforwardly using crude USI and visual pictures (Sec. 2). They portray a few structures (Sec. 2.2) in between multimodal CNN preparing mutually the dual optical modules. As the suggested strategy is compared to our foundation depends on PCA [10] and to a brilliant proposal by which a customary audio-based discourse acknowledgement framework (ASR) prepared on a similar database (Sec. 3). Vitally, we center in this investigation around consistent discourse (instead of detached word) acknowledgment. Results are exhibited and talked about in Sec. 4.

2. OPTICAL PROPERTIES USING CNN

2.1 Convolution neural networks

A Convolutional neural network is basically a group of attributes which can attribute importance to various aspects in an image and can identify the variations among them. The input image is passed through various convolution filters. These filters map the various aspects of the image and each filter clusters certain feature in the signal. Since an image consists if various visual elements, many searches are performed over a single image. In CNN the image is recognized as a three-dimensional object in terms of RGB system rather than as perceived by humans. A CNN architecture reduces the number of parameters as well as reuses the weights which in turn facilitates for easier processing of image. The architectural structure of CNN is made of: 1) a specific number of convolution layers which have been assigned with functions like convolution filtering, 2) a fully connected group of layers which can

be used to learn the non-linear high level features from the output the convolution layer, 3) a softmax layer which performs the task of distinguishing among the high level and low level features so as to classify them.

2.2 Individual analysis of optical modules

In the initial application we utilize dual convolutional neural networks where every processes uni optical modules independently (i.e. video and ultrasound). We imply the slope-descending reverse mobilizing technique in order of analyzing the parameters at the training stage which are provided with phonetic labels as targets. By obtaining the result of the final almost joined layer, we acquire a matrix of optical features for each modality from the connectivity and here we used a multimodal assistant framework in which we can generally process the video and ultrasound pictures. The planning which is a mix of two recognized CNNs is addressed by Fig. 2. It incorporates a mix layer which cements the video and ultrasound modalities. The reason behind this structure is to empty atypical state fuses by in the meantime seeing the upgrades of tongue, lips and the jaw. Subordinate upon this multimodal structure we broke down dual unmistakable strategies for visual highlights extraction: 1) near the finish of the blend layer we separate just a single part vector (Fig.2 top, usage S3), and 2) at the deferred outcome of the final completely related part, simply showing up before the blend layer, we get two segment vectors. The subsequent highlights may differ from the ones got while setting up the two CNNs as it were. These two modalities are bound together and their parameters are assessed together.

2.3 HMM-based optical discourse analysis

This architecture allows former knowledge while decoding by using pronunciation dictionary and language model. In the perspective of optical discourse affirmation such prior knowledge is of great importance. Two methods can be inspected to put together ultrasound and visual modules in the Hidden Markov model Model – Gaussian Mix Model decryptor: 1) an early fusion technique which consisting feature vectors corresponding to every module that can be linked together and demonstrated using a uni-stream HMM-GMM decoder and, 2) a middle composition technique depending on a dual path Hidden Markov Model-Gaussian Markov Model decryptor.

3. BLOCK DIAGRAM

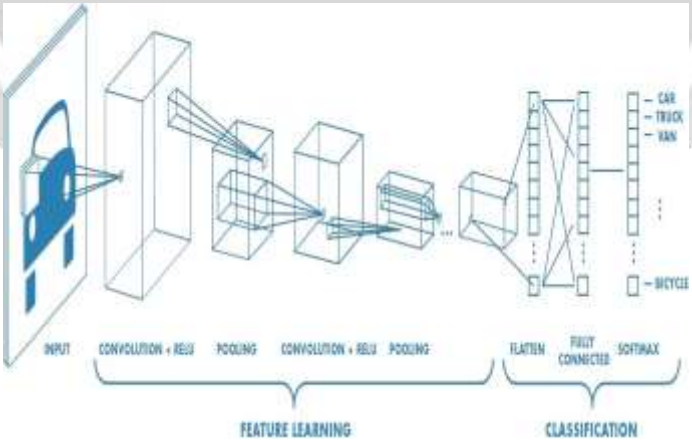

Fig-1: CNN representation

3.1 First Hidden layer

Analytical observations were performed on the database which included 488 sentences which were cleared up by a male French speaker. A Terason T300 pleasing USI structure with a 128 pieces micro convex transducer interruption significance) were used to get ultrasound pictures (320x240 grayscale pictures, 60fps). An advanced CMOS camera was used to record video pictures of speakers face. A change defensive top was used to keep ultrasound and video sensors fixed concerning the speakers face. Ultra-speech forming PC programs was used to record video and sound data in soundproof remain with stable lightning conditions. A great deal of 34 phonemes was used to depict French language. The phonetic transcripts and fundamental motivations behind control of each phoneme were isolated from the recorded verbalizations. The optical information was named by division of sound banner to set up the CNNs and HMM-GMM decoders.

Insistence tests were created without managing before phonetic data, as the present examination concentrates precisely at actuating the limit of the multimodal CNN to process the optical information. Phonemes confirmation accuracy Tp is settled so as to quantify the execution as Np where Np is the measure of phonemes in the test corpus and D, S and I are the autonomously the undoing's, substitutions and consolidations. The phonetic certification's 95% confirmation between time (Δ95%) was figured after [20]. To break our corpus into eight subsets, keeping seven for preparing and one for exploring, 8-spread cross-support is utilized.

The basic CNN structure consists of three approaches- one convolution layer, one full layer and one softmax layer with just a conformist extent of channels acknowledging fulfilling yields. Utilizing another immediate structure: uni complexity layer, uni full layer, uni blend layer and uni softmax layer for frameworks, fulfilling results were gotten. It isn't sure that the proposed structure is immaculate and its tuning will improve. despite whether the free parameters become an important factor. In setting to all CNNs, Rectified Linear Unit(ReLU) non-linearity is utilized for convolution, blend and full layers. Matconv Net apparatus stash is utilized to execute all CNNs which are prepared utilizing GPU-quickening. HTK device compartment with a unique way and a standard arranging system is utilized to build up all HMM-GMM decoders. The imperative subordinates are appeared with their changed visual highlights for all tests. By loosening up the HMM-GMM state back shots utilizing the Viterbi figuring, the no ifs ands or buts game-plan of phonemes is predicted at translating stage. The weighting rules which are utilized to mix the stream probabilities were in addition epitomized on the arranging set for 2-stream HMM structure.

3.2 Fundamentals

The convolutional neural network-based part analysis technique was considered to a PCA-based approach and this system is a slight difference in the Eigen Faces method and goes for searching a reason that best explain the altering of pixel component in as of training graphs. At highlight extraction coordinate, picture whose size is changed and organized video/ultrasound plot is anticipated onto this reason and the visual highlights are given as the D first deals with, for every way. The measure of headings is non-resistant property. In our usage, it is updated on the course of action set by keeping the eigenvectors that pass on 79.9% of the change, which drove for our situation to D = 29 for both mp4 and ultra-volumes pictures. In context on this methodology, we incite two measure structures, where PCA highlights are decrypted utilizing previous mix way or a center blend system.

The execution was additionally separated and the one got while pondering the sound information close-by video, taking into account that sound gives concentrated data, we expect that the ASR definitiveness gives superior way replenished by a VSR structure. Sound flag was parameterized utilizing disintegrating. The HMM-GMM deen crypter was prepared utilizing an equivalent technique in regards to the VSR frameworks.

4. CONCLUSION

We researched the utilization of CNN for extracting visual highlights from ultra-volumes and pictures of the tongue and lips. We investigate a use multimodal design in which the dual optical modalities are together handled. These

two modalities are collaborated together to obtain an effective visual speech recognition system. We inferred diverse frameworks in which the CNN is utilized as a component analyzer and is joined with a HMM-GMM decoder. Tests were led on a ceaseless discourse VSR task. Observation was shown the capability of the CNN over a recently distributed standard. Such tests will be directed in future examinations on a multispeaker database. Future work will likewise concentrate on the structure of a start to finish VSR framework in accordance with present work on ASR joining convolutional layers for handling the crude visual information with an intermittent design to show the elements of discourse explanation The CNN based approaches beats the PCA based baselines paying little regard to the technique used, which executes the limit of the CNNs to expel fundamental properties from the terrible video and ultrasound pictures and the abilities saw between the CNN supported by VSR systems are more hard to introspect , for any condition the running with end is obtained.

5. RESULTS

The output is given by comparing the training inputs and the test inputs and the class is decided. The data that is compared is based on the spectrum parameters like intensity, Amplitude, time period. The data is processed and is presented in the form of wave form. During the test and train phase additional windows for the system is shown to mark the completion of the train functions. The outputs that are got in the training part is the filtered input signal, processed wavelets signal, data matrix set of the processed signal.

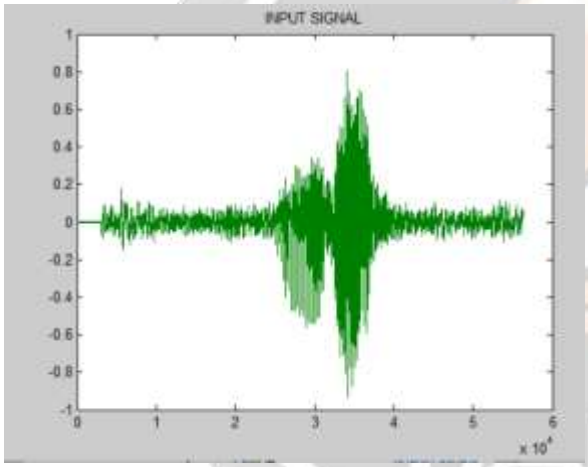

Figure 2: Input signal

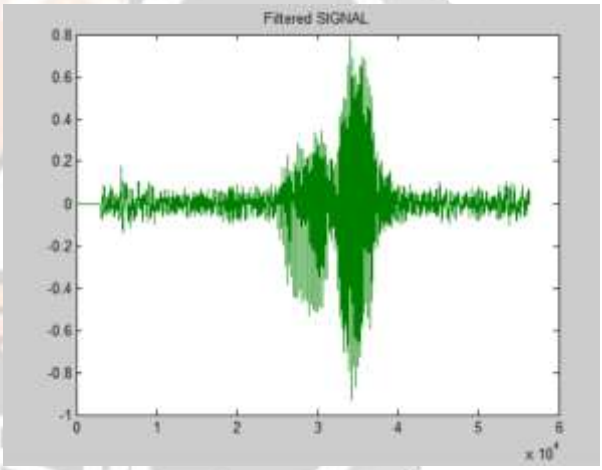

Figure 3: Filtered signal

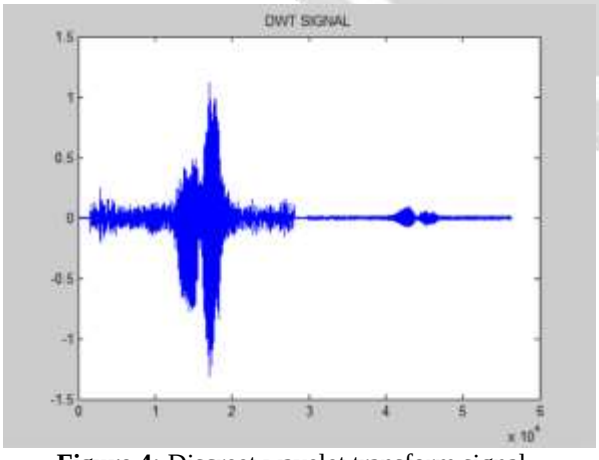

Figure 4: Discreet wavelet transform signal

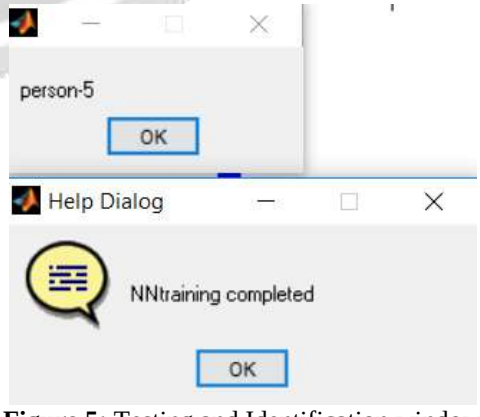

Figure 5: Testing and Identification window

7. FUTURE ENHANCEMENTS

According to the new advances in technology, it has become essential for us humans to embrace the changes in them and apply them in real time. One of the greatest achievements towards this goal is the visual speech recognition technology where one can recognize an individual and communicate in the natural form i.e. vocal cords and speech. Corporations are now utilizing VSR system so that it can recognize customer intentions and desires by inspecting patterns in the voice, stops between the words and the conjugations of these spoken words. This implies that someone talking to a call center bot can be fully understood and the organization can collect full information about customer satisfaction and can provide elite quality service. Besides this, the technology has applications in the field of military services for danger assessment. Also some courts are using this technology where there is shortage of staff. Police departments and US troops in Iraq implement this technology to identify the speech and translate the sentence in the required language. It is also used to give instructions to students where there is shortage of teachers. An important advance can be made to help the physically impaired people so that they can communicate by the VSR system. This technology is currently implemented by Japan and US in robotic android project for facial recognition and mirroring. These robotic bots can be utilized for threat assessment at airports by replacing human workers. It can also be utilized in the case of border crossing by identifying individuals and alerting the concerned authorities. According to the recent advances in US military, vocal cords can be read without exploiting speech or voice, soldiers can communicate with each other. This can be done by fixing a device near the larynx which can detect the vibrations and interpret the speech. The other soldiers can hear this speech by using a tiny ear piece which is silent to others nearby. This can be mainly utilized by the SWAT teams and Special Forces.

8.REFERENCES

[1]. Astik Biswas, PK Sahu, and Mahesh Chandra, "Multiple cameras audio visual speech recognition using active appearance model visual features in car environment," International Journal of Speech Technology, vol. 19, no. 1, pp. 159–171, 2016.
[2]. Thomas Hueber and Gerard Bailly, "Statistical conversion of silent articulation into audible speech using full covariance HMM," Computer Speech & Language, vol. 36, pp. 274–293, 2016.
[3]. A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for MATLAB," in Proc. ACM Multimedia, 2015, pp. 689–692.
[4]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, May 2015.
[5]. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proc. IEEE-CVPR, 2014, pp. 1725–1732.
[6]. Karen Simonyan and Andrew Zisserman, "Two stream convolutional networks for action recognition in videos," in Proc. NIPS, 2014, pp. 568–576.
[7]. Richard Bowden, Stephen Cox, Richard Harvey, Yuxuan Lan, Eng-Jon Ong, Gari Owen, and Barry-John Theobald, "Recent developments in automated lipreading," in Proc. SPIE, 2013, pp. 89010J–89010J–13.
[8]. Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification.," in Proc. BMVC, 2012, pp. 1–12.
[9]. Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg, "Silent speech interfaces," Speech Communication, vol. 52, no. 4, pp. 270–287, 2010.