ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

LIP READING USING NEURAL NETWORK AND DEEP LEARNING

[1]**K. Jyoshna Priya**, Assistant Professor, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
[2]**V.Padmaja**, Assistant Professor, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
[3]**V. Geetanjali Lakshmi**, UG Student, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
[4]**K. Sri Lakshmi**, UG Student, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
[5]**SK. John Basha**, UG Student, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India
[6]**V. Neeraja**, UG Student, Department of Computer Science and Engineering, DVR & Dr. HS MIC College of Technology, Kanchikacherla, Andhra Pradesh, India

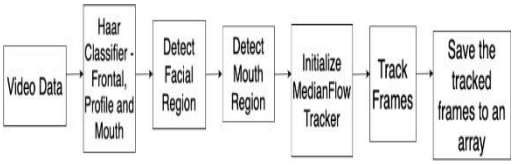
Abstract
Lip reading is a technique to understand words or speech by visual interpretation of face, mouth, and lip movement without the involvement of audio. This task is difficult as people use different dictions and various ways to articulate a speech. This project verifies the use of machine learning by applying deep learning and neural networks to devise an automated lip-reading system. A sub-set of the dataset was trained on two separate CNN architectures. The trained lip reading models were evaluated based on their accuracy to predict words. The best performing model was implemented in a web application for real-time word prediction.
Keywords: Speech recognition, end to end, CNN, LSTM, automated lipreading, and computer

1. Introduction

The interaction quality of contemporary computer vision-based assistance technologies is significantly improved by the ability to use a natural-to-human method ofcommunication. The most common form of human communication is speech. The accuracy and robustness of automatic speech recognition (ASR) systems, on the other hand, are unsatisfactory in many real-worldusage scenarios, such as when driving a car or in a busy area. In these circumstances, the advantage of using visual information about speech (lip-movements) in addition to audio is undeniable and is incorporated in a variety of state-of-the-art systems.

We attempted to use computer vision and machine learning to tackle the issue of automated audio-visual speech recognition in the current study. We created two separate integral (end-to-end) systems employingCNN- based deep neural network architectures for the automatedrecognitionofRussianspeechwitha restricted vocabulary. Also, by training the networks with pictures of speechspectrograms, we attempted to see the challenge of acoustic voice recognition as a purely computer vision issue. For the Russian language, hardly any study has been done in this area. Researchers don't believe there is a ready-made option for creatingsuchsystems.FortrainingNNmodels, therearenorepresentativeopen-accessdatasets that meet the necessary criteria, such as having enough speakers, phone-viseme labelling, a task-appropriate vocabulary size, etc. (Almost no public datasets are available for languages other than English). These elements working together enable us to identify a sizableresearch need.

This study's major objective is to improve automated speech recognition systems' ability to accurately recognise speech in loud environments, which is a crucial challenge.



Lip Detection Process

DOI: 10.36893.JK.2023.V13I04N16.0085-0089

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

2. RelatedWork

2.1 NeuralNetwork

A computer system called a neural network (NN) is analogous in some ways to the nerve system in human brains. The neural network is a framework of algorithms that collaborate to find the fundamental relationships in a dataset in order to deliver the best results. Between the input and output levels, there are so-called hidden layers that combine their individual functions to carry out certain tasks. Hammerstrom specified the use of neural networks in computer systems for a variety of tasks, including image identification, computer vision, character recognition, stock market forecasting, medical applications, and image compression. I utilise a particular kind of deep learning neural network called a convolutional neural network for my lip reading system.

2.2 Convolutional NeuralNetwork

A type of neural network system in a typical multi-layered network is called a convolutional neural network (CNN). The layers are made up of one or more layers joined together in a sequence. With high dimensional data sets like those made up of photos and videos, CNN is able to make use of the local connection. This capability enables CNN to be used in the areas of voice recognition and computer vision.

Convolutional layer, activation function, pooling layer, and fully linked layer make up a basic CNN's four importantcomponents

.Convolutionallayerlearnstheparametersfrom the input data using a set of learningfilters.

2.3 Data &Preprocessing

There aren't many publicly available audio- visual datasets of Russian speech that are good for NN training. The most current one, which wasdebutedinthework,wascreatedespecially for the task of reliable voice recognition in a loud automobilesetting.

A continuous Russian speech with multi-angle videoandaudiodataispartofthemulti-speaker audio-visual corpus known as RUSAVIC (RUSSIAN Audio-Visual speech In Vehicles). 20nativeRussianspeakersmaybeheardonthe recordings.Russianspeechiscapturedonaudio and video in the database, along with labelling data. The software package built to collect, synchronise, and merge audio and video data from two or more cellphones situated in the vehicle cabin was used to record and classify audio-visual data. The speech corpus was recordedinbothtrafficandwhenacarwasidle, i.e.,infull-scaleandsemi-naturalsituationsthat were as similar as feasible to the actual operatingconditions.

Each speaker conducted ten recordingsessions, which were recorded by three cellphones using FullHD 1920 1080 video resolution and a 60 framespersecondframeratefromthreedistinct viewpoints. According to open source data from numerous cutting-edge voice recognition engines, like AlexaAuto, YandexDrive, GoogleDrive, etc., the speaker said 50 of the most common driver requests for cellphones throughout each recordingsession.

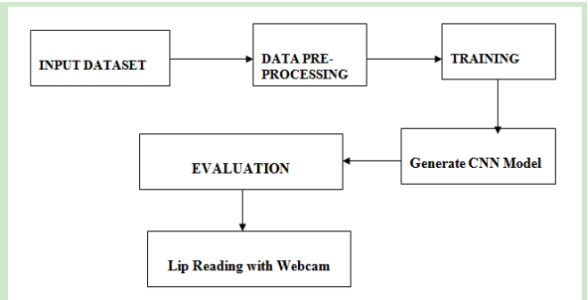


Fig1: Data Flow Diagram

TheDFDisreferredtoasabubblechart.Itisa straightforward graphical formalism that may be used to depict a system in terms of the data that is input into the system, the different processingoperationsthatareperformedonthis data, and the output data produced by this system. The data flow diagram (DFD) is a crucialmodellingtool.Itisutilisedtomodelthe system's parts. These elements include the system's internal workings, the data that supports those workings, an outside party that engages with the system, and the information flows inside the system. DFD demonstrates how data flow through the system and how itis altered by a number of transformations. DFDis

DOI: 10.36893.JK.2023.V13I04N16.0085-0089

sometimes referred to as a bubble chart. Any degree of abstraction may be utilised to portray a system using a DFD. Depending on the amount of information flow and functional intricacy, DFD may be divided intotiers.

2.4 Training

Using a CNN architecture, the lip-reading model was trained. The pre-processed lip samples were trained using a 3-Dimensional (3D) CNN, and various parameters were compared. because of the 3D CNN architecture's ability to train on high- dimensional data, such as image sequences. To train the model, 3D CNN could be loaded with the pre-processed image data that was saved from the training set. Two distinct architectures for 3D CNN models (EF-3 architecture and Lightweight Model architecture). Using Keras, a Python-based open-source neural network toolkit, both topologies were put into practise. A reasonably straightforward neural network implementation is combined with high-performance computationalmethodsinKeras.Similartothat, IranontopoftheTensorFlowframeworkusing the Keras package. A python-friendly open- source framework called TensorFlow offers dataflow programming and numerical computation for deep learning. It enables the deployment of computing across several platforms, including Tensor ProcessingUnit, Central Processor Unit, and Graphics Processing Unit (GPU) (TPU). The CNN modelswerecreatedusingKerasandabackend based onTensorFlow.

3. ProposedMethod

With an end-to-end method to automatic speech recognition, all the steps of the conventional technique are combined during training of a single neural network. This also assumes the availability of certain network structural building components, which we classify into four sequential processing stages:

1. The inputs, which are either a series of cropped mouth photos for lipreading or spectrogram images for acoustic speech recognition.
2. A front-end module that uses the inputs to extract features. For the lip-reading system's visual feature extraction, we employed three to four 3D CNN layers, and for acoustic speech recognition, we used a number of pre- trained CNNs.
3. To model the temporal dependencyand condensethefeaturesintoasinglefeature,usethe back-end module.
4. To calculate the probability of each sentence, use the classification module. represented by a softmax layer in both systems.This structure fits the majority of the existing end-to-end systems. We concentrated on the inputs in this article (visual and audio datapreprocessing)

3.1 Three-Dimensional CNN Algorithm

Step1: On this page
Step2: Setup
Step3: Load and preprocess video data
Step4: Create the model
Step5: Train the model. Visualize the results.
Step6: Evaluate the model.
Step7: Next step

4. Results & Discussion

Theassessmentofthemodelisacrucialtaskin the field of machine learning. To prevent overfittingonthelipreadingmodel,itiscrucial to understand if the trained model has acquired patterns to generalise the prediction in unseen data. For the test set of the LRW dataset, I ran a Top-1 accuracy test to evaluate the predictive accuracy of the model. As a result, the expected responseisthewordwiththehighestlikelihood.

 DOI: 10.36893.JK.2023.V13I04N16.0085-0089

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

Fig2: Haar classifier-detecting mouth region

4.1 Kernel

I increased the kernel size in 3D convolutional layers in Model A2 to (5x5x5) and compared theoutputtothatofModelA1withtheoriginal kernel size (3x3x3). The validation accuracy significantly increased thanks to the larger kernel, going from 66.56% to 70.29%. Here is presented a comparison of the kernel sizes between Model A1 andA2.

Model	Test Accuracy	Kernel Size	Epochs	Speed
EF-3	70.92%	3x3x3	16	190 ms
A1	66.56%	3x3x3	14	104 ms
A2	70.29%	5x5x5	14	110 ms
B	71.25%	5x5x5	17	115 ms
C	74.37%	5x5x5	12	120 ms
D	77.14%	5x5x5	25	117 ms

Fig3: Evaluation of model architecture

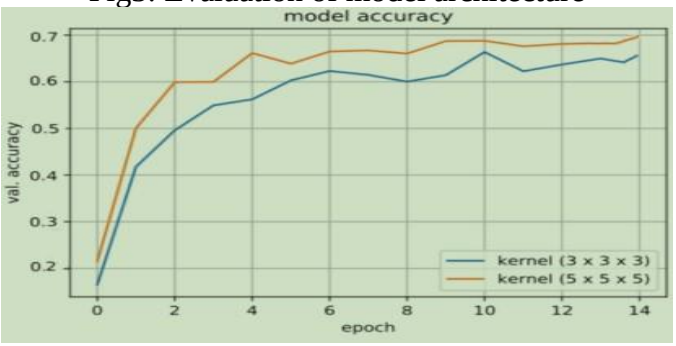


Fig4: Models A1 and A2 are contrasted in terms of kernel size

Conclusion

The main goal of this research is to make speech understandable to those with hearing impairments without the need for specific instruction or human assistance. Lip reading is a way of translating speech to writing without audio input. Hearing impairment is a common problem in society. The effectiveness of computerised lip reading systems is increased when image processing and deep learning are integrated. The accuracy of feature extraction is increased by combining depth maps with 2D photos. We have discussed the history of gesture language implementation, how difficult it is for the average person to use it, and the necessity of lip reading. The fundamental objective of this research is to enable hearing- impaired individuals to comprehend speech without the use of special training or human help. Without audio input, lip reading is a methodofturningspeechintotext.Hearingloss is a widespread issue in society. When image processing and deep learning are combined, computerised lip reading systems perform better.Bymergingdepthmapswith2Dpictures, feature extraction accuracy is improved. We have spoken about how gesture language was first used, how difficult it is for the typical person to utilise, and how lip reading is essential.

References

A. Kashevnik et al.: Multimodal Corp.us Design for Audio-Visual Speech Recognitionin Vehicle

DOI: 10.36893.JK.2023.V13I04N16.0085-0089

Cabin. In IEEE Access, vol. 9(2021)34986-35003.doi: 10.1109/ACCESS.2021.3062752.
A.,Howard,M.Zhu,B.Chen,etal.:Mobile Nets: Efficient Convolutional NeuralNetworks for Mobile Vision Applications. In arXiv:1704.04861, pp. 1-9 (2017). arXiv:1704.04861.
L.Chen.2016.keras,js.https://github.com/tra nscranial/keras-js
[4]D.Ivanko,A.Karpov,D.Ryumin,etal.: Using a high-speed video Camera for robust audio-visual speech recognition in acoustically noisy conditions. In International Conference on Speech and Computer, (2017) 757-766. doi: 10.1007/978-3-319-66429-3_76.
François Chollet. 2015. Keras documentation. *keras. io*(2015).
Yiting Li, Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2016. Lip reading using a dynamic feature of lip images and convolutional neural networks. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. IEEE,1–6.
Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87– 103.

 DOI: 10.36893.JK.2023.V13I04N16.0085-0089