Lips Reading Using 3D Convolution and LSTM

Rohan Inamdar *, Kavin sundarr , Deepen Khandelwal , Ajeyprasaath KB *

Posted Date: 13 December 2023

doi: 10.20944/preprints202312.0928.v1

Keywords: deep learning; computer vision; 3D convolution; lstm; lip reading

Article

Lips Reading Using 3D Convolution and LSTM

Rohan Inamdar *, Kavin Sundarr, Deepen Khandelwal and Ajeyprasaath KB *

SENSE, Vellore Institute of Technology, Chennai, India; kavinsundarr.s2021@vitstudent.ac.in; deependheeraj.khandelwal2021@vitstudent.ac.in

* rohan.inamdar2021@vitstudent.ac.in; ajeyprasaath.kb@vit.ac.in

Abstract: This paper introduces an innovative approach to lipreading, leveraging a web application designed to generate subtitles for videos where the speaker's mouth is visible and a comprehensive literature review that precedes the discussion, encompassing a thorough examination of various lipreading methods employed over the past decade. Our method employs a powerful deep learning model, featuring a 3D-convolution network and bidirectional LSTM, enabling accurate sentence-level predictions based solely on visual lip movements. With an impressive accuracy of 97%, our model is trained using pre-segmented lips regions, transformed into animated GIFs for effective pre-training. This work stands as a significant contribution to the evolving landscape of lipreading research, offering a practical and accurate solution for real-world applications.

Keywords: deep learning; computer vision; 3D convolution; LSTM; lip reading

1. Introduction

Lips reading is a rare talent that some people have and it requires years of experience to improve his or her accuracy and there is a higher chance that a person with decades of experience to make mistakes in lip reading than a trained lip reading model created using deep learning algorithms. As it is very tough to learn and there are not many people practicing it. So, creating a lip reading model with high accuracy does not have any drawbacks and it can be implemented in so many so many fields and applications.

Lip reading has many applications and in the future, it could be a basis for many future developments for example combination of lipreading and audio transcription to auto-create subtitles for any movie or video, Robots in the future could use lipreading and facial emotion scanning to analyze human behavior, by using live lipreading model to convert the predicted text to audio in minimum time delay will give people who can't speak a voice of there own.

In the past when research on lip reading was conducted they used to divide the problem into two parts learning of visual features and prediction based on them. Later the approach was end-to-end trainable and worked only for word classification [4] and the current modified model is useable for end-to-end sentence level, and sequence level prediction [2]. As previous lipreading models created have high accuracy (over 95%) and the model architecture created is good and by training more data the accuracy will increase.

The main goal of this research is divided into two parts first is to train more datasets to increase the accuracy of our prediction and the second part is to create a web application where the user can upload any videos and they will get the predicted text results of the lip movements of the speaker in the video. This can used to create subtitles for any video the user has uploaded This can help people with impaired hearing who want to watch content but don't have subtitles available for that video or the user can convert the predicted text to voice so that people who can't speak would not require to use any hand signs and instead speak normally and attach the audio track created to there video to reach more people.

2. Litrature Survey

In 2016 Assael YM, Shillingford B, Whiteson S and De Freitas N worked on Lipnet: End-to-end sentence-level lipreading. [2] a model trained end-to-end that uses spatio-temporal convolutions, a

 doi:10.20944/preprints202312.0928.v1

recurrent network, and the connection temporal classification (CTC) loss to convert a variable-length stream of video frames into text. This lipreading model simultaneously learns spatio-temporal visual features and a sequence model. LipNet achieves 95.2% accuracy at the sentence level on the GRID corpus [16]. In the end, it was concluded that the performance would improve with more data in the future.

In 2016 Wand M, Koutník J and Schmidhuber J tried using long short term memory(LSTM) for lip reading [4] yielding significantly better accuracy than conventional methods. Feedforward and recurrent neural network layers. The layers are layered to create a single structure that is trained by back-propagating error gradients via each layer. Utilizing common computer vision features, an experimental evaluation and comparison of the performance of such a stacked network against a typical Support Vector Machine classifier was conducted Data from 19 speakers of the publicly accessible GRID corpus [16] were used for the evaluation. It reported the best word accuracy on held-out assessment speakers of 79.6% when employing the end-to-end neural network-based solution with 51 different words to classify.

In 2016 Garg A, Noyola J and Bagadia S used CNN and LSTM to create a lip reading model. They used a VGGNet that has been trained on celebrity human faces from IMDB and Google Images [6] and investigated several approaches for using it to manage these image sequences. The VGGNet is used with LSTMs to extract temporal information and is trained on images concatenated from numerous frames in each sequence. The concatenated image model that uses nearest-neighbor interpolation performed well, obtaining a validation accuracy of 76% while the LSTM models failed to beat other methods for a variety of reasons.

In 2017 Stafylakis T and Tzimiropoulos G [5] suggest a word-level visual speech recognition architecture based on end-to-end deep learning. Bidirectional Long Short-Term Memory, residual, and spatiotemporal convolutional networks are all combined in the system. The network, which consists of a 3D convolutional front-end, a ResNet, and an LSTM-based back-end, was trained using an aggregated per time step loss. This network delivered 83.0% work accuracy, which is 6.8% greater than the 76.2% accuracy of the previous best attentional encoder-decoder network and has an error rate that is less than half that of the reference VGG-M network.

In 2018 Xu K, Li D, Cassimatis N and Wang X [15] produced a deep neural network-based end-to-end lipreading system called LCANet. Utilizing a layered 3D convolutional neural network (CNN), LCANet encodes the input video frames. The roadway system and a GRU network that is two-way. Both short-term and long-term spatiotemporal information is successfully captured by the encoder. What's more, LCANet uses a cascaded attention-CTC decoder to produce output sentences. The results reveal that the suggested system outperforms state-of-the-art techniques by 12.3% on the GRID corpus database with a 1.3% CER and 3.0% WER.

In 2018 Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G and Pantic M [7] presented an end-to-end audiovisual model based on residual networks and Bidirectional Gated Recurrent Units (BGRUs). Under clean and low noise conditions, the end-to-end audiovisual model slightly outperforms a conventional MFCC-based system. Additionally, it performs noticeably better than both the end-to-end and MFCC-based audio models when there is a lot of noise. This system was only able to recognize isolated words it was not capable enough to recognize sentences.

In 2022 Ajeyprasaath KB, Vetrivelan P.[19] loking at the the network traffic brought on by multimedia streaming and live streaming in 5G New Radio (NR) creates a number of issues, including matching network operator supply, user expectation, and difficulty in network design and conservation.In order to estimate user expectations and improve QoE performance, they proposed a Supervised Machine Learning classification technique [20]. Using Dynamic Adaptive Streaming over HTTP (DASH) for video and multimedia streaming in a real-time simulation scenario derived from 5G traces in static and mobility instances, the QoE performance is evaluated.

In 2020 Martinez B, Ma P, Petridis S,and Pantic M [9] address the limitations of the model created using residual network and Bidirectional Gated Recurrent Unit (BGRU) layers in the audiovisual model,

 doi:10.20944/preprints202312.0928.v1

suggest adjustments that would enhance its performance even more. First, Temporal Convolutional Networks (TCN) are used in place of the BGRU layers. Second, they significantly streamline the training process, enabling them to train the model in a single step. Thirdly, they demonstrate that models created using the most recent state-of-the-art methods do not generalize well to differences in sequence length. To solve this problem, they suggest a variable-length augmentation. The results for isolated word recognition in English and Mandarin using the largest publicly accessible datasets, LRW and LRW1000, respectively. The absolute improvement produced was 1.2% and 3.2%, respectively.

In 2021 Ma P, Petridis S and Pantic M [8] give a hybrid CTC/Attention model that can be trained end-to-end and is built on a ResNet-18 and Convolution-augmented transformer (Conformer), an encoder-decoder attention-based architecture for audio-visual speech recognition that can be trained end-to-end. The audio-visual model greatly outperforms the audio-only model, especially at high levels of noise, according to results on the two largest publicly available datasets for sentence-level speech recognition, Lip Reading Sentences 2 (LRS2) and Lip Reading Sentences 3 (LRS3), respectively used to test the architecture.

In 2021 Sarhan AM, Elshennawy NM and Ibrahim DM [14] developed a deep neural network for lipreading called hybrid lipreading (HLR-Net). The model has three stages that create the output subtitle: preprocessing, encoder, and decoder. The encoder, which implements connection temporal classification (CTC), is built using the inception, gradient, and bidirectional GRU layers, and the decoder, which implements CTC, is built using the attention, fully connected, activation function layers. On the GRID corpus [16] dataset, the proposed HLR-Net model can significantly outperform the three most recent models, LipNet[2], Lip-reading Model with Cascaded Attention (LCANet), and attention-CTC (A-ACA), achieving CERs of 4.9%, WERs of 9.7%, and Bleu scores of 92% in the case of unseen speakers and CERs of 1.4%, WERs of 3.3%, and 99% in the case of overlapped speakers.

In 2023 Miled M, Messaoud MA and Bouzid A [10] Using a hybrid model with a new proposed edge based on a proposed filter, extracted the mouth region and segmented the mouth, and then trained a spatio-temporal model by combining convolution neural networks (CNN) and bi-directional gated recurrent units (Bi-GRU). The accuracy rating of the algorithm was 90.38%.

3. Methodology

We declared vocabulary for every possible single character that can be expected by using Keras create number to character and character to number function. From the dataset by loading Alignments given specific paths and split outlines. If the first line contains 'sil' (silence) will be ignored and append the rest to the array and use characters to number encoder. Create a Data pipeline to train deep learning model. Tensorflow would draw random samples from the dataset to complete 1 training step. The segmentation of the mouth region is done statically and by using Imageio and its mimsave to create an animation gif that our model will learn to decode.

3.1. Dataset

We have used the GRID corpus[16] dataset to evaluate our model as it is sentence level. It is an extensively used voice corpus created for research and development in the fields of audiovisual speech recognition (AVSR) and automated speech recognition (ASR). The phrases in the GRID corpus are made to cover a variety of English phonetic information. The dataset includes recordings from 34 speakers, including 16 women and 18 men. A wide variety of English dialects are represented by these speakers. A collection of 1,000 phonetically balanced sentences was recorded by each speaker. The audio data provides clean speech recordings with minimal background noise, while the video data captures the speakers' facial movements during speech articulation.

3.2. Model Architecture

Our model is created using 3D convolution and LSTM combination (overview in Figure 2) as it is a powerful approach to process sequential data with both spatial and temporal dependencies.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2023 doi:10.20944/preprints202312.0928.v1

Figure 1. Pre-Processing of Videos.

The 3D convolution Network is better for working with videos. Similar to 2D convolution(spatial convolution), 3D convolution works by moving a kernel (also called a filter) across the input data to extract local characteristics. The kernel's size corresponds to the depth, height, and width of the data, and it spatially moves over the input volume. The kernel calculates the element-wise dot product with the relevant input sub-volume at each point. To create a feature map with high-level representations of the input volume, this process is repeated for all points. The output of the CNN is then fed into the LSTM as sequential data, where the LSTM captures temporal dependencies and patterns. In this architecture dense, dropout, and bidirectional can convert paths through Temporeal components while using LSTM. As for the optimizer, Adam optimizer was used the final model was implemented using streamLit

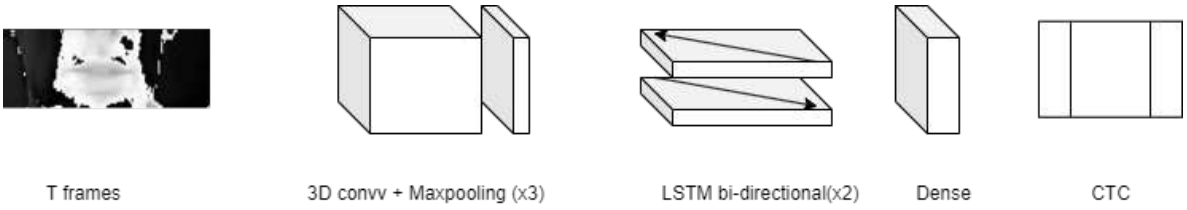

Figure 2. Model Architecture.

The classification dense layer is used to make our predictions. We have used our special loss function CTC(Connection Temporal classifier) as it works better for word transcripts that are not aligned to frames and to avoid and reduce the duplicates.

Layer (type)	Output Shape	Param #
conv3d (Conv3D)	(None, 75, 46, 140, 128)	3584
activation (Activation)	(None, 75, 46, 140, 128)	0
max_pooling3d (MaxPooling3D)	(None, 75, 23, 70, 128)	0
conv3d_1 (Conv3D)	(None, 75, 23, 70, 256)	884992
activation_1 (Activation)	(None, 75, 23, 70, 256)	0
max_pooling3d_1 (MaxPooling3D)	(None, 75, 11, 35, 256)	0
conv3d_2 (Conv3D)	(None, 75, 11, 35, 75)	518475
activation_2 (Activation)	(None, 75, 11, 35, 75)	0
max_pooling3d_2 (MaxPooling3D)	(None, 75, 5, 17, 75)	0
time_distributed (TimeDistributed)	(None, 75, 6375)	0
bidirectional (Bidirectional)	(None, 75, 256)	6660096
dropout (Dropout)	(None, 75, 256)	0
bidirectional_1 (Bidirectional)	(None, 75, 256)	394240
dropout_1 (Dropout)	(None, 75, 256)	0
dense (Dense)	(None, 75, 41)	10537

Total params: 8,471,924
Trainable params: 8,471,924
Non-trainable params: 0

Figure 3. Model layers and size overview.

4. Results and Discussion

After training our model, we noticed whatever input video was provided the output sentences were accurate but our overall performance was a little low (refer to Table 1) compared to other models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2023 doi:10.20944/preprints202312.0928.v1

that used the GRID corpus dataset as we had trained our model with the limed dataset(about over 1000 videos).

Table 1. performance comparison for the GRID corpus dataset only.

Model	Year	Accuracy
Lipnet[2]	2016	95.2
Lipnet(with face cutout)[17]	2020	97.9
LACNet[15]	2018	97.9
CTC/Attention[18]	2022	98.8
Our Model	2023	97

As you can see in Figure 4 the overall website looks like it is an easy to use website that gives out predicted output in text format and also provides the Gif of lip movements that our model used as input to get the predicted output. So in cases where we find wrong output, we can first check if the lip moments are correctly segmented.

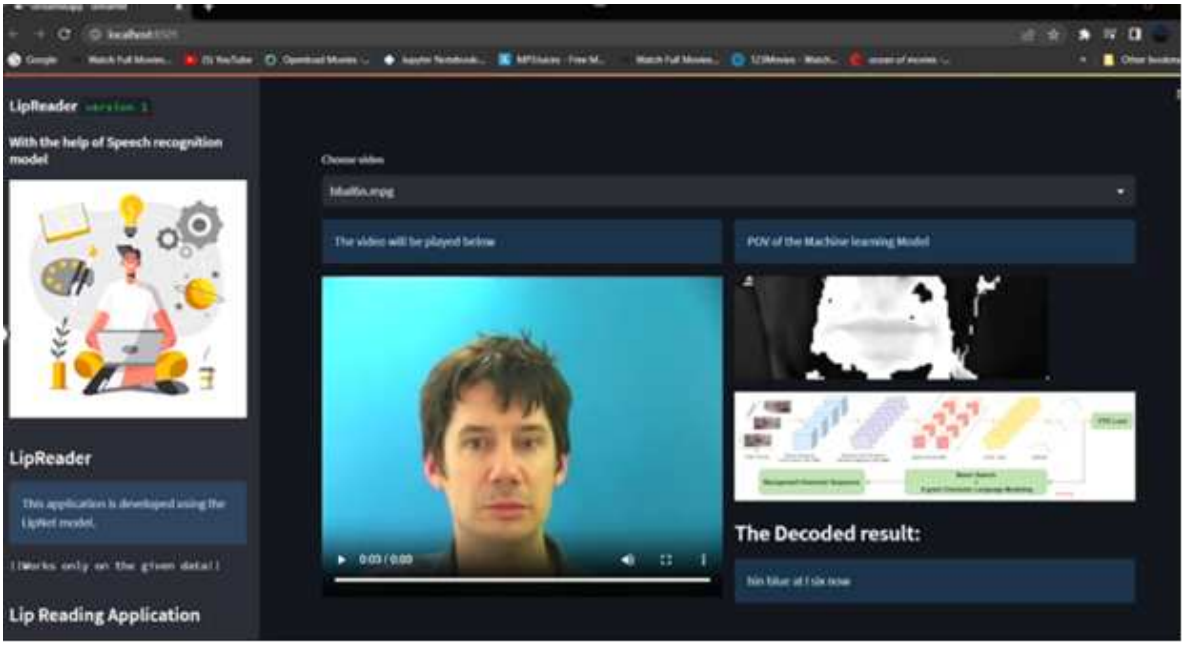

Figure 4. overview of website.

5. Conclusions and Future Scope

We propose a web application LipReader where users can upload any video in which mouth movements are visible in the frame and get the predicted end-to-end sentence level text which can be used as subtitles for the video. The accuracy of our model is 97% . The predicted text can reduce the time for subtitle creators and people with hearing impairments. Also, people who can't speak can record their videos get predicted text, and use a text-to-audio converter to have their own voice instead of using sign language.

For future work, we can create a model jointly trained with audio-visual speech recognition for end-to-end sentence level prediction, where visual will help in predicting the subtitles for videos with noisy backgrounds and also training even larger datasets to get higher performance or use transformers instead of LSTM for state of the art.

References

1. Chen W, Tan X, Xia Y, Qin T, Wang Y, Liu TY. DualLip: A system for joint lip reading and generation. InProceedings of the 28th ACM International Conference on Multimedia 2020 Oct 12 (pp. 1985-1993).
2. Assael YM, Shillingford B, Whiteson S, De Freitas N. Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599. 2016 Nov 5.
3. Garg A, Noyola J, Bagadia S. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report. 2016.
4. Wand M, Koutník J, Schmidhuber J. Lipreading with long short-term memory. In2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 Mar 20 (pp. 6115-6119). IEEE.
5. Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105. 2017 Mar 12.
6. A. V. Omkar M Parkhi and A. Zisserman, "Deep face recognition," Proceedings of the British Machine Vision, vol. 1, no. 3, p. 6, 2015.
7. Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G, Pantic M. End-to-end audiovisual speech recognition. In2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2018 Apr 15 (pp. 6548-6552). IEEE.
8. Ma P, Petridis S, Pantic M. End-to-end audio-visual speech recognition with conformers. InICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021 Jun 6 (pp. 7613-7617). IEEE.
9. Martinez B, Ma P, Petridis S, Pantic M. Lipreading using temporal convolutional networks. InICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020 May 4 (pp. 6319-6323). IEEE.
10. Miled M, Messaoud MA, Bouzid A. Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications. 2023 Jan;82(1):551-71.
11. Manaswi NK, Manaswi NK. Understanding and working with Keras. Deep learning with applications using Python: Chatbots and face, object, and speech recognition with TensorFlow and Keras. 2018:31-43.
12. Bradski G, Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc."; 2008 Sep 24.
13. Howse J. OpenCV computer vision with python. Birmingham: Packt Publishing; 2013 Apr 23.
14. Sarhan AM, Elshennawy NM, Ibrahim DM. HLR-net: a hybrid lip-reading model based on deep convolutional neural networks. Computers, Materials and Continua. 2021 Jan 1;68(2):1531-49.
15. Xu K, Li D, Cassimatis N, Wang X. LCANet: End-to-end lipreading with cascaded attention-CTC. In2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) 2018 May 15 (pp. 548-555). IEEE.
16. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.
17. Zhang Y, Yang S, Xiao J, Shan S, Chen X. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) 2020 Nov 16 (pp. 356-363). IEEE.
18. Ma P, Petridis S, Pantic M. Visual speech recognition for multiple languages in the wild. Nature Machine Intelligence. 2022 Nov;4(11):930-9.
19. Ajeyprasaath KB, Vetrivelan P. A QoE Framework for Video Services in 5G Networks with Supervised Machine Learning Approach. InInternational Conference on Machine Intelligence and Signal Processing 2022 Mar 12 (pp. 661-668). Singapore: Springer Nature Singapore.
20. Ajeyprasaath KB, Vetrivelan P. Machine Learning Based Classifiers for QoE Prediction Framework in Video Streaming over 5G Wireless Networks. CMC-COMPUTERS MATERIALS & CONTINUA. 2023 Jan 1;75(1):1919-39.