International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Lip Reading Using Machine Learning and Neural Networks

Mr. Shivaraj B G[1], Akash[2], Abhishek Ashokan Nambiar[3], Harith Harish[4], Shamir L C[5]

Student, Dept. of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidre, India[2,3,4,5]

Assistant Professor, Dept. of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidre, India[1]

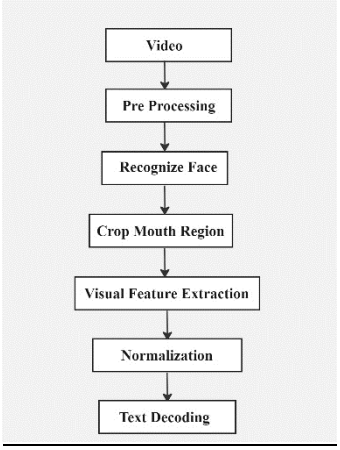
Abstract: Lip reading, often overlooked in communication, involves visually interpreting lip movements to understand spoken words. This process entails recognizing lip positions and movements, organizing them into sound sequences, and decoding sentences. Machine learning can play a crucial role by training models on labeled datasets of lip movements and corresponding phonemes. These models can then extract features from new lip movements to classify spoken phonemes. Such technology can assist individuals with hearing impairments by improving speech recognition in noisy environments and aid security forces in situations lacking audio records. By integrating innovative technologies seamlessly, our solution aims to empower people with hearing loss to engage more fully in society and bolster security measures.

Key Words: Lip Reading, Machine Learning, Convolutional Neural Network (CNN)

1. INTRODUCTION

Lip reading is a skill that allows people to understand speech by observing the movements of the speaker's lips. It is a useful skill for people with hearing loss, as well as for people who need to communicate in noisy environments or with people who speak different languages. This project aims to develop a lip-reading system using machine learning which is designed to be helpful when say there is a situation where we need to extract conversations from recorded footages but audio data is not available or reliable, or a noisy environment, where audio/speech detection becomes unreliable and erroneous, also it will help to improve the quality of life of those people with hearing and speech impairments. The system will be trained on a dataset of lip movements labeled with the corresponding words and phrases. Once trained, the system will be able to recognize words and phrases from new lip movements. The system will be developed using a deep learning model. Deep learning models are a type of machine learning model that are particularly well-suited for tasks such as image recognition and speech recognition. In this context deep learning, image processing is used in lip reading systems because they are particularly well-suited for tasks such as feature extraction and pattern recognition.

2. METHODOLOGY



Preprocessing: Initially, the video undergoes preprocessing by dividing it into frames. These frames, initially in RGB format, are converted to grayscale to streamline processing and reduce parameter overhead. The resulting frames are then subjected to further analysis.

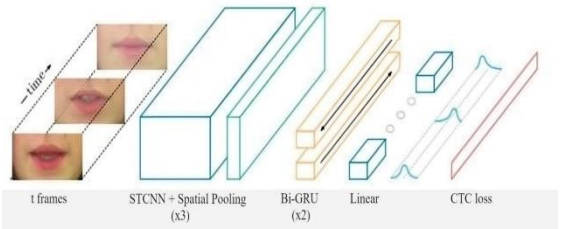
Face Detection and Cropping: Following frame extraction, the system employs a face detection mechanism, specifically targeting full frontal views using DLib's face detector and landmark predictor with 68 landmarks. Frames lacking a detected face are discarded. Subsequently, the system identifies the Region of Interest (ROI), focusing on the lips and mouth area using a Haar cascade classifier. This region is cropped using an affine transformation, resulting in mouth-centered crops of 100*50 pixels. RGB channels are standardized to zero mean and unit variance before saving the cropped images as a NumPy array.

Feature Extraction and Normalization: Features are extracted from the ROI, emphasizing spatio-temporal characteristics, which are then fed into a Convolutional Neural Network (CNN) for training. Normalization is

applied to ensure uniformity in training data, accommodating variations in speaking speed among individuals.

Text Classification and Decoding: Normalized data is supplied to the CNN for training and subsequent text decoding. Through numerous epochs, the CNN autonomously learns and deciphers lip movements, predicting the spoken words. Predicted words are then aggregated to reconstruct the original sentence spoken in the video.

Architecture: The system architecture is based on a CNN framework comprising an input layer, three hidden layers, and an output layer. A SoftMax layer serves as a probability classifier, while max pooling reduces parameter complexity in subsequent layers. The hidden layers are configured with 32, 64, and 96 neurons, respectively. While a 5-layer architecture was considered, computational constraints favor the prioritization of the 3-layer architecture.



Building the model: The model begins with a series of Convolutional Layers, aimed at extracting relevant features from the input data. These convolutional layers are followed by Max Pooling operations to reduce spatial dimensions and maintain important features. After the convolutional layers, a Time Distributed layer is utilized to apply the Flatten operation across the temporal dimension, preparing the data for subsequent processing. Following this, the architecture incorporates two stacked Bidirectional Long Short-Term Memory (Bi-LSTM) layers. These layers are adept at capturing intricate temporal dependencies within the input sequences, owing the bidirectional nature.

To mitigate overfitting and enhance generalization, dropout regularization is applied after each LSTM layer. Finally, the model concludes with a Dense layer, employing the softmax activation function to generate probability distributions over the output classes. Overall, this architecture combines convolutional feature extraction with LSTM-based sequence modelling, rendering it capable of handling complex sequential data with both spatial and temporal characteristics.

The trained model can now predict speech visually by matching the position and movement of lips to corresponding tokens. The result is displayed as text for the given input video.

3. LITERATURE REVIEW

Abiel Gutierrez et al.,[1] present a variety of models and methods for predicting words from video data without audio, the data was pre-processed by using various existing facial recognition software to detect and crop around the subject's faces and these frames were used as input model. Various models were tried for this CNN + LSTM Baseline model, a Deep Layered CNN + LSTM model, an ImageNet Pretrained VGG-16 Features + LSTM model. Inclusion of pre-trained facial recognition CNNs highly improved the models. The fine-tuned VGG + LSTM gave the best test results, with 59% test accuracy compared to CNN + LSTM Baseline model, a Deep Layered CNN + LSTM model. In all models, they found it very difficult to avoid overfitting with unseen people. The model worked better depending on whether it was working with seen or unseen people for testing and validation, high overfitting is observed for seen people. The input sequence of frames were five. Facial recognition was used to identify and crop the entire face, which was fed as input.

In the paper by **Souheil Fenghour et al.,[2],** a neural network-based lip reading system is proposed. The system is lexicon-free and uses purely visual cues. With limited number of visemes as classes to recognize, the system is designed to lip read sentences covering a wide range of vocabulary and to recognize words that may not be included in system training. BBC Lip Reading Sentences 2 (LRS2) benchmark dataset was used. It focuses on improving the accuracy of lip reading sentences rather than any word-based approach by using visemes. A viseme-based lip reading system can be used to classify words that have not been presented in the training phase, and they can be generalized to different languages because many different languages share the same visemes. A CNN-based detector (Single Shot MultiBox Detector), is used for detecting face appearances within the individual frames and to recognize facial landmarks. The classification accuracy of visemes achieved by the proposed system was very high (over 95%), the classification accuracy of words was significantly dropped after the conversion (65.5%). When the visemes are misclassified, they are most likely decoded as one of the most frequently appearing visemes in training data. This leads to an increased number of errors.

Triantafyllos Afouras et al.,[3] have proposed and compared three new neural network architectures for lip reading based on sequence learning methods. Two models use CTC, a recurrent model with LSTMs and a fully conventional model. Another model involves the use of an attention-based methods. BBC-Oxford Lip Reading

Sentences 2 (LRS2) benchmark dataset was used. A decoding algorithm was proposed for online lip reading. The training proceeded in three stages first, the visual front-end module is trained; second, visual features are generated for all the training data using the vision module; third, the sequence processing module is trained. Silent video of a talking face was used for learning, to predict the sentences being spoken. Decoding algorithm was developed for online lip reading, however during real time large number of incomplete words were found. Diminishing returns in terms of accuracy gains when training on sequences longer than 80 frames was observed. The model is less accurate when longer sequences are inputted.

In the paper by **Kuniaki Noda et al.,[4]** a convolutional neural network (CNN) is used as a visual feature extraction mechanism for VSR. By training a CNN with images of a speaker's mouth area in combination with phoneme labels, the CNN acquires multiple convolutional filters, used to extract visual features essential for recognizing phonemes. The proposed system is evaluated on an audio-visual speech dataset comprising 300 words with six different speakers. The evaluation results of the isolated word recognition experiment demonstrate that the visual features acquired by the CNN significantly outperform those acquired by conventional dimensionality compression approaches, including principal component analysis. The experimental results demonstrate that a supervised learning approach to recognize phonemes from raw mouth area image sequences could discriminate 40 phonemes by six speakers at 58% recognition accuracy. The proposed system prepared a trimmed dataset by manually cropping 128×128 pixels of mouth area from the original visual data and resizing them to 32×32 pixels. Manually cropping for a larger dataset is hectic and impractical. There are significant variations in mouth area appearance, depending on the speaker. The model is speaker dependent for phoneme recognition as the dataset is small and insufficient. Introducing new and unseen speakers results in high inaccuracies. The dataset is small and makes the model speaker dependent. A larger dataset would have helped cover all possible appearance variations.

4. SUMMARY

Audio recognition is an important tool used in different segments of the present technological world. However, the sole dependence on audio for speech recognition can result in low accuracy of the deciphered data due to noise as well as unreliable audio. In such circumstances, we feel the need to use an additional technology that can overcome the shortcomings of audio detection. The movement and position of lips, generally used by the hearing impaired to decipher words and sentences, can be implemented using deep learning to assist speech recognition. Hence, lip reading can hugely improve and assist speech recognition technology which can be used in various fields ranging from subtitle generation to security.

5. CONCLUSION

Lip reading using machine learning and neural networks highlights the model which processes video frames and generates text sequences seamlessly through end-to-end training. It prioritizes constructing coherent sentences over individual word deciphering, giving 92% accuracy rating, surpassing the average lip-reading software by 13%.

For future work, several avenues for improvement and expansion are identified. Some of them are:

Multi-language Lip Reading: Extend the model to recognize lip movements in multiple languages, catering to diverse linguistic communities.

Real-time Applications: Optimize the model for real-time performance, enabling its integration into various applications like virtual assistants, video conferencing, and accessibility tools.

Improved Accuracy: Continuously refine the model with larger datasets and advanced CNN architectures to enhance accuracy, especially in challenging environments with varying lighting conditions and speaker characteristics.

Gesture Recognition Integration: Integrate gesture recognition with lip reading to enhance communication accessibility for individuals with hearing impairments.

Cross-Modal Learning: Explore cross-modal learning techniques to incorporate additional cues such as audio and facial expressions to further improve lip reading accuracy and robustness.

REFERENCES

[1] "Lip Reading Word Classification" by Abiel Gutierrez, Zoe-Alanah Robert (2020)

[2] "Lip Reading Sentences Using Deep Learning with Only Visual Cues" by Souheil Fenghour, Daqing Chen, Kun Guo (2020)

[3] "Deep Lip Reading: a comparison of models and an online application" by Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman (2018)

[4] "Lipreading using Convolutional Neural Network" by Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakada

[5] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in NIPS, 2012.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning forimage recognition," arXiv preprint arXiv:1512.03385, 2015.

[8] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. ICML, 2015.

[9] Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R., "Extraction of visual features for lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198–213, 2002. https://doi.org/10.1109/34.982900

[10] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H.Glotin, D. Vergyri, J. Sison, Mashari, and J. Zhou, "Audio-visual speech recognition," Center Lang.Speech Process., Johns Hopkins Univ., Baltimore,MD, 2000. https://doi.org/10.1016/j.imavis.2014.06.004

[11] K. Cho, B. van Merrienboer, C¸ . G ¨ ulc¸ehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in EMNLP, Oct 2014.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine trans-lation by jointly learning to align and translate," CoRR, vol.abs/1409.0473, 2014. 1