Scope
Volume 14 Number 02 June 2024

Lip Reading Using Connectionist Temporal Classification

Mala B M; Meghana K; Adhira M Nair; Sparsha B; Lekhana M

School of Computer Science
Reva University
Bengaluru, India

Abstract: Lip-reading is the responsibility of decoding text from the movement of a speaker's mouth. Lip-reading system takes video without audio as an input of a speaker speaking some word or phrase and provides the anticipated word or phrase as output. This is exceedingly beneficial for hearing impaired individuals to understand the movement of the mouth of a speaker who do not know sign language in the physical world with a lot of noise pollution. Conventional methods have concentrated mostly on bulk preprocessing. Regardless of showing immense potential, application of deep learning algorithms has been minimal in this field. Here we expose a Convolutional Neural Network (CNN) model to anticipate words from video without the audio. Lip-reading system also uses an attention-based Long Short-Term Memory (LSTM), Connectionist Temporal Classification (CTC) along with Convolution neural network (CNN). The trained lip-reading model is evaluated based on the accuracy to anticipate words. Moreover, we examine challenges and limitations associated with deep learning-based lip reading, including data scarcity, variations in lighting conditions, speaker-dependent variability, and occlusions. To address these limitations and improve system performance, we propose the adoption of ensemble learning techniques in future iterations. This research contributes to the advancement of lip-reading technology, particularly beneficial for hearing-impaired individuals navigating noisy environments where sign language is impractical. By harnessing deep learning methods, we aim to enhance accuracy and efficiency, thereby improving accessibility and communication for diverse populations.

Keywords: Preprocessing, Deep Learning, Convolution Neural Network, Long Short Term Memory

I. Introduction

Speech recognition from the movement of the lips is generally used in noisy or loud environments. This is predominantly helpful for people with hearing disabilities.

Parallelly, for security purposes, this lip-reading system can be used to determine and predict the information from the speaker when the audio is absent or corrupted in the video. The ability to recognize the text from only the visual information is a magnificent skill, but a challenging task to any beginner. This task is challenging as people use different dictions and several ways to articulate a speech. There are wide variety of languages spoken around the world with the large difference in diction and relative articulation of words and phrases. It has become a challenging task to design a computer program that accurately recognizes the words and phrases from the movement of the speaker's mouth alone without audio. Even the expert lip-readers can only be able to find about every second word. Consequently, with the help of neural networks and deep learning algorithms, it is made possible to automatically and accurately determine the text from the movement of the speaker's mouth without audio. Moreover, the integration of multimodal approaches, combining visual lip movements with audio signals and contextual information, offers opportunities for even greater accuracy and robustness in lip reading systems. Fusion techniques that leverage both visual and auditory modalities can mitigate the limitations of each modality individually, resulting in more reliable speech recognition outcomes. Additionally, the integration of domain knowledge from linguistics, phonetics, and cognitive science into deep learning-based lip-reading models can enhance their interpretability and linguistic accuracy. By incorporating insights from these disciplines, researchers can design more linguistically informed models capable of capturing subtle linguistic features and improving word recognition accuracy. By addressing challenges, leveraging multimodal information, integrating domain knowledge, and exploring novel methodologies, researchers can drive forward the state-of-the-art in lip reading and unlock new opportunities for enhancing communication, accessibility, and security.

II. Related Works

In this part, we describe different ways to use computers to read lips.

A. Automated Lip Reading

Most lipreading technology does not use deep learning. Working on videos involves either processing the frames to find image features, processing the frames over time to find video features (like detecting movement), or using other methods. Handmade systems of seeing things. Papandreou and his team wrote some papers in 2007 and 2009. Pitsikalis and his team also wrote a paper in 2006. Lucey and Sridharan wrote a paper together in 2006. Papandreou and his team also wrote another paper in 2009. There are many books and articles about automated lipreading. In 1997, a group of people used technology to read lips in videos without sound for the first time. They used a method called hidden Markov models and had a small amount of data to work with. They also manually separated the spoken sounds in the videos. Later, Neti and others in 2000, they were the first to understand speech by watching someone's mouth move and using a statistical model called HMM. Using features that were created by hand, on the IBM ViaVoice. The writers make it easier for computers to understand speech in loud places by combining what they hear with what they see on screen. The dataset has 17111 spoken sentences from 261 speakers for practice. As mentioned, their results that only show visuals cannot be understood as only visual. Understanding, because they are used to improve the quality of the unclear audio recordings. We can do the same thing using a similar method. Potamianos and his colleagues in 2003, a report showed that the accuracy for recognizing speech from any speaker was 91. 62%, and for recognizing speech from a specific speaker it was 82. 31%. The connected DIGIT corpus had a word error rate of 38. 53% use the same dataset. This contains numbers written as words. Moreover, Gergen and others trained the speaker on a transformed version of the LDA in 2016. The process of converting mouth movements into data using a mathematical method in a speech recognition system. This job has the best performance in the GRID corpus, with a speaker-specific accuracy of 86. 4%. It's still a challenge to understand how to use motion features and apply them to different people.

B. Classification with Deep Learning

This involves using advanced computer algorithms to categorize data into different groups or classes based on their characteristics or features. These techniques have shown promising results in various applications, such as medical diagnosis, image recognition, and natural language processing. However, all these ways only work on individual words or sounds, but this project can predict whole sentences. One way to do it is by studying how we understand information using both sound and sight. Researchers like Ngiam, Sui, and Ninomiya have investigated this method. In 2015 and 2016, Petridis and Pantic learned how to recognize visual features along with speech processing. Almajai used different methods to group words and sounds together into categories. Many of these methods are like the early use of neural networks to process speech in speech recognition. Chung & Zisserman (2016a) suggest using special types of neural networks for working with spatial and spatiotemporal data. Using the Visual Geometry Group (VGG) model to classify words. The designs are checked using a dataset that looks at individual words. BBC TV (333 and 500 classes) hasn't been as good at predicting where and when things happen as other models. It's been off by about 14%. Also, their models cannot work with different values. Different lengths of sequences and they don't try to predict entire sentences. In 2016, Chung and Zisserman trained a model that matches audio and visual cues to learn mouth features. They used these features as inputs to classify 10 different phrases using a Long Short-Term Memory (LSTM) algorithm. OuluVS2 dataset and a task that doesn't involve reading lips. Wand and others in 2016, a new type of neural network called LSTM was used for lipreading. However, it did not cover certain issues. Predicting the order of sentences or determining the speaker's identity are not factors in this case. Garg and others in

2016, a VGG model that was already trained to recognize faces was used to identify and sort words and phrases. MIRACL-VC1 dataset has 10 words and 10 phrases in it. But the thing that happens repeatedly is that it is the best for them the model is trained by keeping the VGGNet parameters the same and only training the Recurrent Neural Network(RNN), instead of training both at the same time. Their top model can only correctly classify 56. 0% of words and 44. 5%. The accuracy of grouping things into categories is the same for both tasks, even though they both involve sorting things into 10 different groups.

C. Sequence prediction in speech recognition

Speech recognition is when a computer can understand and predict what someone is saying, changing the way we live and work. The world today would be very different without new technology that helps us learn deeply. It has changed how we do things in our daily lives and at work. In Automatic Speech Recognition (ASR) text needs to be rewritten in a simpler way to make it more understandable. The Connectionist Temporal Classification loss (CTC) was created by Graves and his team. In 2006, the movement was driven. Deep learning is used as a part of ASR, to create ASR systems that are trained from start to finish. There has been a lot of improvement in lipreading recently. The progress in ASR has been like earlier progress, but it has not reached sequence prediction. Our project is a model that can predict whole sentences by looking at lip movements. Basically, the project uses a series of images as its input. It creates a list of tokens in a certain order. It is trained all at once using CTC.

D. Deep Learning for Speech Recognition

Speech recognition technology has improved a lot. Mainly because deep learning research is doing well and the ability to understand and repeat what someone has said. The number of mistakes is already lower than that of a person. Giving commands through speech provides a simple method for instructing the computer, portable electronic devices or integrated technology. Techniques for comprehending natural language have been improved through the use of advanced methods. Newer computer programs and equipment are moving fast. The intelligent virtual assistant has methods of providing assistance. Virtual assistants like Siri, Alexa, Echo, Google Now, and Samsung Bixby have become popular products. The best speech recognition models are very successful. The model was built using either the CTC or sequence-to-sequence approach. Methods based on CTC utilize predictive algorithms. Each sound part is given a specific name and is trained to improve prediction accuracy by tweaking its alignment, and the label we are trying to reach. Seq2seq models help in understanding and recognizing speech. "Got a lot better with focus mechanism" which helps make information flow better than before. A type of neural network that learns patterns in data over time. Attention-based sequence-to-sequence model has done better than the CTC based model in many comparisons. Data sets are collections of data that are stored in a specific way for easy access and analysis. New research demonstrates that focus and concentration are important. The model might not work well with speech data that has a lot of background noise that suggests a shared way to figure out spoken words in a full speech recognition system using a mix of CTC and attention structure. The mixed method reduces an overall loss by combining the losses of both parties into one. CTC loss and attention loss are used together in a way that works well. Benefits of understanding and interpreting something. LCANet is different because it uses a different approach. A decoder that pays attention to different parts of the input and uses a method called CTC. The advantage of this type of design is that this helps to avoid adjusting the weight parameter in the program. The loss function is responsible for directly generating the output, rather than combining the results from two sources. The softmax output is utilized instead of merging both CTC and focus divisions. Instead of aggregating the outcomes, the loss function generates the output directly.

E. Machine Lipreading

Understanding what is being said in simple terms can be challenging, but let's give it a try. Machine lipreading usually involves two main steps: the visual part, where the machine "reads" the lips, and the understanding part, where it interprets the words being spoken. Analyzing how the lips move and guessing what the next 4 text will be, using categorizers Traditional features that are seen visually are taken out. 1) The area around the mouth can be divided into four main categories based on the visual features that are directly extracted using DCT (Discrete Cosine Transform) to change images DWT; 2) features based on shapes, for example, height, the width of the lip area; 3) features based on motion, for example, Motion of the lips seen through a camera; 4) features based on a model. The way the lips look and their shape are called the lip ROI. ASMs are utilized to accurately depict. Active Appearance Models (AAMs) are a type of technology used to analyze and track the appearance of objects in images. The quality of these model-based features performance depends greatly on their level of excellence. The correctness of the training data that is labeled by hand which needs a lot of effort working diligently to attain something. The alterations occurring in visual characteristics are utilized, while the static classifier is replaced with dynamic classifier such as hidden Markov models (HMMs). However, these classifiers are built on the condition of independent assumption that doesn't work well for predicting long-term relationships. Recently, people are using deep neural networks, Comprehending the movements of the lips and excelling when compared to the typical methods, suggesting using CNN for identifying and collecting visual aspects, as well as detecting a mix of Gaussian observations. Creating a system for recognizing single words by studying how the sounds change over time. Identify the order of labels. suggests a complete sentence-level understanding of what people are saying by looking at their lips. A 3D CNN that is stacked on top of each other is used for moving images. Extracting visual characteristics from images. The connection between how the lips move and the words they represent is shown by the Gated Recurrent Unit (GRU). CTC loss is a method that is used to figure out the difference between the guesses and the text shows labels that are different lengths. The system provides a full lipreading function that processes two input streams, which are static visuals containing information from the mouth. Variations exist in how people communicate during different periods in the same location. RBMs are used to initially train the encoding layers and LSTM models track the evolution of things over time. The coming together or combination of the two streams flow back and forth through a Bi-directional path. LSTM (BLSTM) suggest using both visual and audio information. The qualities of spoken language can be utilized to identify speech in diverse manners. Both video and audio are used as sources of input. Characteristics of speech are employed in various methods to recognize speech. This method uses the 'EESEN' framework as its basis. Uses CTC loss function to match the timing. Issue The second method is based on the idea of observing and paying attention.

F. Deep Learning Video Super-Resolution(VSR)

In the past few years, deep learning techniques have been used successfully in many different areas, including VSR systems. Deep learning methods are better at making accurate predictions than old-fashioned ways. For example, when a CNN is used together with traditional methods, the trained classifier CNN design can tell the difference between visemes. Then an HMM framework is added to give time information after the CNN gives its output. Other scientists used LSTM and HoGs together to analyze short phrases from the GRID dataset. In the same way, a computer program was used to make guesses about words by analyzing information from the OuluVS and AVLetters datasets. The sequence-to-sequence model (seq2seq) is a type of deep speech recognition system that reads the input and then predicts the output sentence. This model uses information from all around the world for longer sequences. Watch, listen, attend, and spell (WLAS) was the first model to use both audio and visual parts to understand speech in videos. It was used to recognize speech in a real dataset. LipNet was the first model to predict letters by reading lips at the sentence level. This model used different techniques to analyze both the space and time aspects of the data and was taught using a particular method to minimize errors. We used a small set of words and grammar rules to test how well LipNet works. The results showed that LipNet made mistakes on 4. 8% of words in some cases, and 11. 4% in other cases. Meanwhile, humans understood 47. 7% of the words in the same test. Similar designs have been created to study how sound and visual features come together. They used a small set of 18 speech sounds and 11 words to predict digit sequences. They used a CTC cascading model for this. This means that the deep learning method can understand and find important features in the experimental data, and it works well with a lot of data and unclear images.

III. METHODOLOGY

Lip reading is the art of understanding the lip movements and predicting the sentence from it. The project is an end-to-end sentence predicting model which is a neural network architecture. It involves collecting the video datasets, frame extraction, lip detection, cropping the lip to the desired size, and then predicting the sentence.

A novel approach in lip reading involves seamlessly mapping variable-length sequences of video frames to corresponding text sequences. This method is trained end-to-end, meaning it learns directly from the data to accomplish this mapping task. The different steps are represented below.

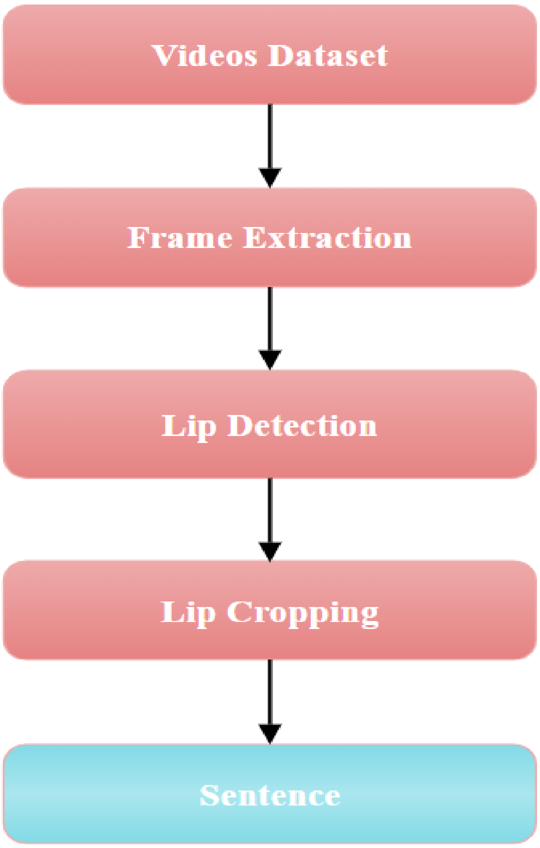

Fig. 1. Architecture Diagram

A. Video dataset

The initial step in developing a lip reading system is to amass a substantial dataset comprising video and audio recordings of individuals engaged in speech. This dataset serves as the foundation for training the system to recognize and interpret lip movements accurately. To ensure the dataset's quality and consistency, preprocessing techniques are applied to address various challenges inherent in real-world recordings. This preprocessing stage encompasses several essential tasks, such as filtering out background noise and irrelevant visual information, normalizing the data to 6 standardize factors like resolution and frame rate, and aligning the audio and visual components to synchronize lip movements with corresponding speech. These techniques help enhance the dataset's clarity and relevance, providing a clean and coherent input for subsequent stages of model development.

B. Frame Extraction

Frame extraction plays a pivotal role in lip reading, as it involves isolating key frames from video recordings that capture significant lip movements during speech. This process is crucial for generating a dataset conducive to training machine learning models. Frame extraction typically involves techniques such as face detection and tracking, where algorithms identify and follow facial features, particularly the mouth region, across successive frames. After looping through each frame and storing them in an array named "frames," the next step involves reducing these frames to distill the essential visual information they contain. This reduction process aims to extract meaningful features from the frames, which can then be used to decode the speech depicted in the lip movements. Frame extraction is essential for creating a focused and relevant dataset that facilitates accurate lip reading model training.

C. Lip Detection

In lip reading techniques, lip detection plays a fundamental role in isolating the region of interest (ROI) containing the lips from the rest of the face and background in a video frame. Lip detection algorithms are designed to identify and localize the lips accurately, despite variations in factors such as lighting conditions, facial expressions, and camera angles. Once the lips are successfully detected, the extracted lip region can be further processed for feature extraction and analysis in subsequent stages of the lip reading pipeline. Accurate lip detection is crucial for ensuring the quality and reliability of the visual input used for lip reading, ultimately contributing to the overall performance of the lip reading system.

D. Lip Cropping

The process involves iterating through each video, wherein the frames are stored in arrays. The frames undergo conversion from RGB to grayscale, a step that reduces the data volume for preprocessing purposes. Subsequently, the mouth region is isolated using a predefined slicing function applied statically. This isolation enables focused analysis on the lip movements. Following this, standardization takes place, involving the calculation of mean and standard deviation. This practice aims to scale the data appropriately. The standardized data is then cast to float32 and divided by the standard deviation, ensuring consistency and normalization within the dataset. Overall, this methodical approach ensures efficient processing and normalization of the visual input, optimizing it for subsequent stages of analysis in the lip reading system.

E. Sentence

The 3D convolutions are used to pass the videos and condense it down to a classification dense layer that predicts single characters. Then the loss function is used, which is known as CTC, Connectionist Temporal Classification to handle the output. The Connectionist Temporal Classification (CTC) loss, introduced by Graves et al. in 2006 and widely used in modern speech recognition systems, is valuable because it eliminates the need for training data to align inputs with target outputs. This advancement, cited by Amodei et al. in 2015 and Graves & Jaitly in 2014, as well as Maas et al. in 2015, is instrumental in addressing the challenge of variable-length sequences. CTC operates by computing the probability of a sequence while marginalizing it over all equivalent sequences, effectively removing the necessity for alignments. It works with a model that generates a series of discrete distributions over token classes, including a special "blank" token. This method ensures that the probability of a sequence is defined without explicitly aligning inputs to outputs. Let's denote the set of tokens classified by the model at a single time-step as V (vocabulary), and the augmented vocabulary as

$$V^{\sim} = V \cup \{-\} \qquad (1)$$

where the blank symbol is denoted by {-}. The function

$$B: V^{\sim} * \rightarrow V* \qquad (2)$$

is defined to remove adjacent duplicate characters and blank tokens from a string over $V^{\sim}$. For a label sequence $y \in V*$, CTC calculates the probability $p(y|x)$ by summing over all sequences

$$u \in B\text{–}1(y) \qquad (3)$$

such that the length of u is equal to the number of time-steps T in the sequence model. For example, if T = 3, CTC computes the probability of the string "am" as the sum of probabilities of "aam," "amm," " am," "a m," and "am ". This computation is efficiently performed using dynamic programming, enabling maximum likelihood estimation.

Overall, CTC's methodology facilitates effective training of speech recognition models without requiring explicit alignments, making it a powerful tool in modern speech processing tasks. The loss function and the callback are found out to check the progress of the model. The CTC loss is defined. The model is tested using a video to see if it gives a correct prediction. The predicted sentence is displayed.

IV. Results

The advancement of deep learning models marks a substantial leap in lip reading accuracy when contrasted with traditional methods. With abundant tagged data at our disposal, deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have surpassed previous benchmarks, attaining state-of-the-art performance in lip reading tasks. These models exhibit remarkable robustness against diverse challenges such as lighting variations, speaker variability, ambient noise, and other confounding factors. By leveraging large datasets, deep learning frameworks excel at discerning intricate patterns, enabling them to effectively account for variations in lip movements induced by different conditions.

Consequently, deep learning has emerged as a powerful tool for enhancing the accuracy and reliability of lip reading model systems, opening new avenues for applications.

The below table shows the accuracy of the different projects
which have used different algorithms.

Table I. Results Of Different Algorithm

Methodology	Accuracy
VGGNet along with SVM	76%
VGG16 and LSTM	59%
Viseme concatenation and 3D CNN	76%
CTC and LSTM	86%

V. Conclusion

To sum up, this paper has provided an in-depth look at the promising field of lip reading with deep learning. Through our research, we have shown that the use of deep learning algorithms such as CNNs and RNNs can be effective in lip reading tasks. We have also discussed how our research can be applied in various applications, such as assistive technologies for hearing impaired, human to human interaction, and surveillance. Looking forward, the lip-reading field using deep learning is well-positioned for further progress. In the future, we may explore multimodal approaches that integrate audio and visual cues to improve recognition accuracy. Ethical considerations such as privacy and bias mitigation will also be important as these technologies are used in real-world environments.

References

[1] A. Zisserman and J. S. Chung. In the wild, lipreading. In the 2016a Asian Conference on Computer VisionA Treatise on Electricity and Magnetism by J. Clerk Maxwell, Third Edition, Volume 2. Clarendon Press, Oxford, 1892, pp. 68–73.

[2] A. Zisserman and J. S. Chung. Automated lip sync in the wild is out of time. In ACCV, 2016b.K. Elissa, "Title of paper if known," unpublished; Workshop on Multi-view Lip Reading.

[3] X. Li, D. Hu, and others. Audiovisual speech recognition through temporal multimodal learning. pp. 3574–3582, IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[4] S. Cox, R. Harvey, J. A. Bangham, T. F. Cootes, and I. Matthews. identifying and removing visual cues for lipreading. IEEE Transactions on Machine Intelligence and Pattern Analysis, 24(2): 198–213, 2002M. Young, The Handbook for Technical Writers. University Science, Mill Valley, CA, 1989.

[5] C. Neti, D. Vergyri, J. Sison, P. Potamianos, J. Luettin, I. Matthews, H. Glotin, and A. Mashari. audiovisual speech recognition. IDIAP Technical Report 2000

[6] Pitsikalis, V., Papandreou, G., and Katsamanis, A. Audiovisual speech recognition using multimodal fusion and learning with uncertain features used. In Multimedia Signal Processing Workshop, pp. 266–267, 2007.

[7] P. Maragos, V. Pitsikalis, A. Katsamanis, and G. Papandreou. Audiovisual speech recognition using adaptive multimodal fusion with uncertainty compensation. 2009; 17(3): 423–435, IEEE Transactions on Audio, Speech, and Language Processing.

[8] M. Pantic and Petrides. Visual speech recognition using deep complementary bottleneck characteristics. On pages 2304–2308 of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2016.

[9] S. Cox, R. Harvey, Y. Lan, and Almamajai. Deep neural networks and speaker adaptive training provide improved speaker-independent lip reading. In the IEEE International Conference on Speech, Signal Processing, and Acoustics, pp.2722–2726, 2016.

[10] N. Jaitly and Graves. Towards recurrent neural networks for end-to-end voice recognition. pp. 1764–1772, International Conference on Machine Learning, 2014.

[11] J. Schmidhuber and A. Graves. Phoneme categorization in frames using bidirectional long short-term memory and alternative neural network structures. In 2005, Neural Networks, 18(5): 602–610.

[12] J. Schmidhuber, A. Graves, S. Fernandez, and F. Gomez. Recurrent neural networks are used to label unsegmented sequence data in connectionist temporal classification. In ICML, 2006, pp. 369-376