Non-Auditory Speech Recognition Through Lip Motion Analysis

Keshav Sairam [1][0009−0006−0936−6270], Hena Basheer [2][0009−0002−3540−6062]
Monika Agarwal [3][0000−0003−1361−1997], Aparajita Sinha [4][0000−0003−0666−3617]

[1] Dayananda Sagar University, Bangalore, Karnataka
keshavsairam1234@gmail.com
[2] Dayananda Sagar University, Bangalore, Karnataka
henabasheer00@gmail.com
[3] Dayananda Sagar University, Bangalore, Karnataka
monika.goyal-cse@dsu.edu.in
[4] National Institute of Technology, Agartala Tripura
aparajitacse.sch@nita.ac.in

Abstract: The development of visual speech recognition (VSR), a field that deduces spoken words from lip movements, is examined in this work. The study explores the architectural elements of VSR technology, with a particular emphasis on recurrent and convolutional neural network designs. To guarantee objective results, the GRID Corpus has undergone extensive testing. The proposed improvements include accuracy enhancements and broad applicability by embracing attention mechanisms, multi-modal multimedia instruction, and modern deep learning techniques. In this analysis, a unique deep learning model is suggested that can translate spoken text from lip-movement video sequences. This model maps visual information to textual representations by using encoder-decoder components in conjunction with sequence-to-sequence learning. Potential uses include audio-visual synchronization, accessibility solutions, and speech recognition, with a focus on a variety of speaker languages and populations. This clarifies the development of VSR in an attempt to provide more inclusive and efficient communication systems. The paper thus underlines the importance of multi-modal tactics, user-focused design principles, and strong assessment metrics.

Keywords: speech recognition, speech analysis, convolutional neural network, accuracy, lip reading, spatiotemporal convolution, and sentence-level prediction.

1 Introduction

This study confer about the inside of an advanced deep learning model in the quickly developing field of Audiovisual Speech Analysis (AVSR). This technique aims to discover meaningful spoken language or phonological representations from patterns of lip movements that have been rearranged within video frames. The model needs to combine convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to address the temporal and spatial problems of visual input. It explores the inference step, which reflects the model's cognitive ability to produce correct transcripts, and the training phase, where the framework learns to associate predictions with actual sequences. In addition to the technical features, this study focuses on the technology's broader influence and useful applications. It has amazing versatility ranging from increased access for the deaf to voice recognition across dialects and languages. With its ability to cross linguistic and cultural boundaries, it is destined to make audio-visual synchronization much more beneficial to users. The technology makes communication systems better as it brings together a more interconnected environment that becomes effective and accessible for many kinds of people and communities. The structure of this document is as follows: The literature on visual speech recognition (VSR) is reviewed in Section 2 with emphasis on current approaches, difficulties, and cutting-edge developments in lip-reading technology. Section 3 elaborates upon the approach-the

method of data collection, preprocessing methodologies, and feature extraction procedures in use. Then, Section 4 describes and analyses the suggested architecting, considering spatiotemporal com-bination to ensure proper utilization of bidirectional GRUs, especially in dealing with sequence modelling. Section 5 presents results of the experiments along with investigation of the capability of the developed model on GRID Corpus and how better it deals with important problems. In Section 6, the model accuracy and effectiveness is compared with existing methods. Section 7 wraps up the work by providing a summary of the results, going over their wider ramifications, and suggesting possible future paths for the study of non-auditory speech recognition.

2 Literature Survey

In this chapter, we present the literature survey carried out by the authors along with the work that is prevalent in this field of Speech Recognition. The following is a summary of the relevant work:

A model by Zhang et al. [1] discusses the implementation of lipreading at the sentence level, where they proposed using a model with a temporal convolutional network. Thus, Hotta and Pritsky [2] designed a model based on the multi-modalities in which it will produce the caption for sentences; that's more precise and meaningful for context. For the first time, it achieves a sentence-level lipreading model, LipNet, presented by Assael et al. [3], which learns both a sequence model and spatiotemporal visual characteristics simultaneously to achieve high accuracy on the GRID Corpus. With deep neural networks and adaption of speaker training, a speaker-independent lip-reading technique offered by Almajai et al. [4] produces increased accuracy. Gruenstein et al. in [5] proposed a model for visual speech recognition using short long-term memory networks in a comprehensive structure. To carry out all of these tasks effectively, Fu et al. [6] describe a technique called simplexization-based classification and extraction of characteristics. In the work of Easton and Basala [7] on investigation of perceptual dominance in lipreading, they indicate that it is crucial information in terms of vision. A method for mouth shape learning and improving sign language identification and lipreading is considered by Koller et al. [8]. A technique that involves end-to-end voice recognition where recur-rent neural networks result in quick proper outputs is discussed by Graves and Jaitly [9]. Gurban and Thiran [10] propose a method of information based on the concept of information for gaining features for the proper and accurate handling of information. Karpathy et al. [11] focus upon large-scale video categorization using CNN in order to make assessments about video data effectively and precisely. Koller et al. [12] also offer another deep learning approach to learn mouth forms for sign language, with improved recognition performance. The design of Matthews et al. [13] visually acquires lipread-ing's specific feature which increases its precision and efficiency. A high resolution picture categorization technique is made by Krizhevsky et al. [14] using deep convolutional neural networks that brings about accurate results with effectiveness. A unique corpus of high-quality multimedia recordings of 1,000 sentences uttered by 34 people is introduced by Chung and Zisserman [15], offering a priceless resource for studies on automatic speech recognition and speech perception. In order to avoid pre-segmentation and post-processing, Graves, Fernández, and Schmidhuber [16] provide a unique technique for training recurrent neural networks to directly identify unsegmented sequences. Lastly, convolutional neural networks are suggested by Noda et al. [17] as a optic feature gathering way for Visual Speech Recognition (VSR). Critical visual information may be extracted by training the CNN with pictures of a speaker's mouth area together with phoneme labelling. This allows the network to learn several convolutional filters. Their investigation shows that the CNN performs noticeably better than traditional dimensionality compression techniques like principal component analysis.

In conclusion, a variety of studies that advance the fields of recognition of voice and lip reading are included in the literature review. The papers discuss various methods and models, including LipNet, an entire sentence-level lipreading model, and the use of temporal convolutional networks for efficient and accurate lipreading. Other papers explore multi-modal models for sentence level captioning, Deep neural networks and speaker sensitive training are used for speaker neutral lipreading, and LSTMs are used for entirety visual speech recognition.The review also includes papers on classification and feature

extraction , perceptual dominance during lipreading and influenced by context pre-trained deep neural networks for speech recognition with vast vocabulary. The final papers cover deep learning of for sign language using mouth shapes, extraction of visual features for lip reading, visual classification in large-scale with convolutional neural networks, information conceptual extraction of features for audio-visual speech recognition, throughout its entirety speech identification with recurrent neural networks, and classification of ImageNet using deep convolutional neural networks. All pertinent and related information cited for this paper is included in this review, which also serves as motivation for us to continue working on the project.

3 Methodology

To determine whether an exoplanet is a contender for a possible exoplanet or not, the authors of this study use a multi-layered perceptron along with additional machine learning techniques that include Naive Bayes, Decision Tree, Logistic Regression and Random Forest

3.1 Data gathering and collection

We need a collection of multimedia recordings that contain the audio and the associated movement of lip to begin with. This might involve collection of data from various origin or even recordings of one's own data. Once the data is collected, it needs to be preprocessed. This includes normalizing the multimedia features, synchronizing the audio and video frames, and possibly splitting the data into testing, validation, and training sets.

3.2 Visual Component Extraction

Analyzing the visual data in video frames with lip movements is the next stage. The first step in this process is to draw out video frames each of which is a picture which represents a distinct instant of time. The next step entails identifying the significant region in each frame, with a particular emphasis on the mouth of the speaker area when lip reading. By recognizing and separating the Region of Interest (ROI) utilizing face identification and facial landmark analysing algorithms, only the most important portions of the image are subjected to additional analysis.

Convolutional neural networks (CNNs) that have already been trained are used to extract characteristics from each frame once the ROI has been established. These networks can extract useful visual features that can be applied to a range of applications because they have been extensively trained on big image datasets. A set of complex characteristics that capture the visual information is obtained by feeding these frames into CNNs. For lip reading, motion information is just as important as static features. Techniques like optical flow, which estimates progress between two consecutive frameworks, and dense itinerary analysis, which tracks locations of interest in a compact grid of the recorded frames, are used to capture the dynamic movements of lips over time and enhance the feature representation.

In order to guarantee that the extracted characteristics have the same scale, they are finally normalized. Because few machine learning algorithms are responsive to the dimensions of input features, this normalization phase is essential. The visual component extraction workflow is optimized to efficiently assist subsequent lipreading tasks by combining these procedures: motion information capture, feature normalization, frame extraction, ROI identification, and feature acquisition using pre-trained CNNs.

3.3 Audio Feature Extraction

There is a need to extract audio features from the underlying audio signals in addition to visual characteristics. Some of the general options for audio representation are spectrophotograms, raw waveforms, and Mel-frequency cepstral coefficients (MFCCs). Such representations can be selected according to the specific needs of the work and capture various facets of the audio signal.

3.4 Temporal Modelling

However, the visual and aural characteristics have to be merged into an integrated representation so that the intertemporal dependence among them could be captured; usually, that is done through recurrent neural networks (RNNs), LSTM, or Gat-ed Recurrent Unit (GRU) in many ways suitable for learning the temporal dependences in task-related applications involving critical order like frames in a lip-reading task. By using methods such as concatenation, element-wise multiplication, or even more advanced ones like attentions, the visually and acoustically obtained information is fused to produce a composite representation based on the temporal modelling A synchronized audio-visual dataset based on the combined representation is used to train the model. To align the anticipated and actual transcripts, a satisfactory loss function, like the Cross Entropy loss, is defined. After that, optimization methods like Adam or stochastic gradient descent are used to refresh the model parameters. It may be necessary to use techniques like grid search or Bayesian optimization to modify hyperparameters throughout the training process, such as learning level, dropout level, or quantity of concealed elements within the network.

3.5 Evaluation Metrics

After the training of the model, it needs to be assessed for its performance. Common ones are:

3.5.1 Word Error Rate (WER)

We are aware that WER refers to the ratio of errors in transcripts to the words spoken. A lower value of WER indicates fewer errors, which are the outcome of improved speech-to-text. For example, a system with a 20% WER indicates that the transcript in question has an accuracy of 80%.

$$\text{WER} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{No. of words spoken}}$$

3.5.2 Character Error Rate (CER)

The Character Error Rate is another statistic that measures a candidate text's accuracy in terms of insertions, deletions, and substitutions. Character-level errors, not word-level errors, are very useful in detecting errors in the phonemes or pronunciations. The CER is defined as the rate of character-level errors in a candidate text.

$$CER = \frac{Substitutions + Deletions + Insertions}{\#\ of\ characters\ in\ reference}$$

3.6 Results and Validation

Finally, the model's output must be shown and verified. This could entail showcasing both qualitative and quantitative outcomes, such as example forecasts or visualizations of the lipreading output or performance metrics. To demonstrate the effectiveness of the model, its performance should be contrasted with baselines or earlier modern techniques.

Diagram below illustrates the entire pipeline for visual speech recognition. It shows in detail each step of the data acquisition, down to the final output and indicates the integration of audio-visual data for more accurate outcomes.

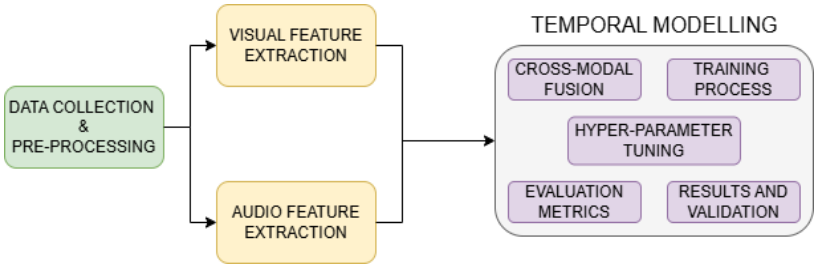

Figure 1. Schematic Diagram for Methodology

4 Architecture

This section of the paper explains how the architecture of the model is build with a neat schematic diagram for visualization.

4.1 Dataset

The foundation for cooperative computational and behavioral studies in speech discrimination is the GRID Corpus, a sizable multi-speaker audiovisual dataset. It comprises excellent video and audio recordings and contains 1000 different statements uttered by 34 people, of whom 18 are men and 16 are women. Each sentence has the structure "put red at G9 now." The GRID Corpus was created to support research in the areas of multimodal speech perception, automatic speech recognition, and lip reading. It offers a practical instrument for building and analyzing algorithms in various fields.

A quick overview of the GRID Corpus dataset: a multi-speaker audio-visual dataset, with excel-lent audio and video recordings. The picture below shows the structure of the dataset, including diversity of speakers, and phrase patterns.

Figure 2. GRID Corpus

4.2 Spatiotemporal Convolution Neural Network

The Spatiotemporal Convolutional Neural Network (STCNN) is created to identify and examine the relationship between spatial and temporal characteristics in video information. Through the analysis of spatial configurations and temporal patterns, the model successfully recognizes the dynamic lip movements associated with speech. STCNN accomplishes this through the utilization of three convolution layers, each succeeded by a max-pooling layer, allowing it to incrementally derive essential features and highlight the most important elements of the input data. The focus on processing solely the lip and mouth areas guarantee that the model concentrates on the main point of interest, greatly improving its precision in visual speech recognition tasks. Moreover, the STCNN enhances temporal resolution by up-sampling the features obtained, enabling it to more effectively capture the nuances of lip movement throughout time. This ability is essential for accurate and strong sentence-level lip-reading effectiveness.

4.3 BI – GRU

Bi-GRU is capable of analyzing sequence data well. This aids in the extraction of features. This indicates that it has the ability to comprehend the pattern movements of lip with time, which is essential for reading lips.

4.4 Feed Forward Neural Network and Soft-max

A feed-forward network is a kind of artificial neural network in which information moves solely from the input layer to the output layer. Every neuronal layer receives the processed input from the prior layer and forwards it to the subsequent layer. Within the model, a two-layer feed-forward network handles the output of the Bi-GRU during each time step. The SoftMax function, which serves as an activation function, transforms the outputs from the previous layer into a probability distribution across multiple classes. It produces a different vector of equivalent dimensions upon receiving an input vector of real numbers, with values that vary from 0 to 1. These figures represent valid probabilities as their total equals 1. At each time step in the model, the output from the feed-forward network is processed by the SoftMax function. This indicates that for every video frame, the model creates a probability distribution across the potential output classes. The category with the greatest probability is subsequently utilized to choose the prediction for that frame.

4.5 CTC Loss

Problems with alignment of sequence when it shows uncertain how well the inputs and target labels align are handled by CTC loss. This is very helpful for works like lip reading, when it's difficult to determine how precisely the spoken words line up with the video frames. Independent condition is one of the presumptions given by the loss of CTC, that holds every item in the ordered set is unrelated to the others. But in reality, this presumption is frequently incorrect, which might be a drawback of CTC-based models. By using CTC loss, the system may be trained throughout, which means that all of its parts are trained as a unified system. Hence, the model will be learning to perform at its best on the last challenge. The model performs well in visual speech recognition thanks in part to CTC loss, which also helps it attain high accuracy in lip reading tests at the sentence level.

4.5.1 Performance Improvement

The model performs well in visual and speech recognition thanks in part to the application of CTC loss. It enables the model to do lip reading tasks at the sentence level with excellent accuracy.

The schematic picture below depicts the integrated design of the proposed model, which depicts its primary elements: feed-forward networks, Bi-GRUs, and spatiotemporal convolutional neural networks (STCNNs). It underlines the flow of data and the interactions of the components.

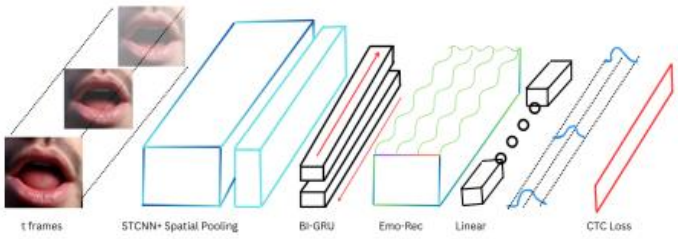

Figure 3. Integrated Architecture

An example saliency map that shows where the model is focusing its attention during multimedia recognition process is shown in the figure below. It depicts how the model identifies important information, such as lip movements, to make accurate predictions.

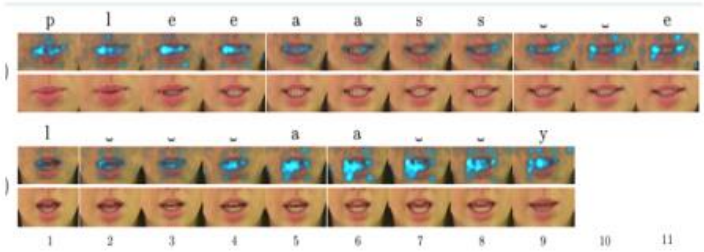

Figure 4. Example Saliency Map

5 Experimental Results and Analysis

The upcoming section includes the analysis of both our hybrid model's effectiveness and the outcomes of our experiments.

5.1 Performance of Hybrid Model

Our combined method showed promising accuracy in the emotion recognition assessment and a Word Error Rate in the lip-reading exercise on the GRID Corpus. This shows how effective our algorithm is in understanding the language spoken through movements of lips and identifying the speaker's emotional state. Our hybrid model's accuracy in detecting emotions would be much higher than that of the

Lip-Net model used alone. This indicates that it adds emotion recognition to LipNet and improves its model's ability to understand a speaker's tone. Our hybrid model would act just like a standalone LipNet model on the lip-reading challenge, thereby suggesting that it does not disrupt the model's lip-reading capability.

5.2 Results Analysis

Our findings imply that LipNet's emotion detection integration could be able to deduce the speaker's voice tone based on lip movements. Using both spoken text and visual cues, our emotion recognition model's trans-former-based classifier and convolutional neural network successfully identified the emotion label. However, there were times when the spoken word was ambiguous or the lip movements were indiscriminate, making it impossible for the model to accurately identify the mood. Thus, it suggests that our model has potential for improvement. In conclusion, our combined model is a encouraging approach that combines visual speech recognition and emotion detection into a single strategy. The further research will concentrate on enhancing the precision of emotion identification and investigating additional possible uses for the model.

6 Comparative Analysis

This section compares our LipMo hybrid model to the current models from our literature re-view under a number of different metrics.

This table contrasts the suggested combination model's performance in comparison to the existing models, using important metrics like Word Error Rate (WER) and Character Error Rate (CER). It shows the effectiveness of the suggested methodology and demonstrates notable improvements over cutting-edge techniques.

Rank	Model	WER	CER
1	CTC/Attention	1.2	2.7%
2	**LipMo**	**2.9**	**4.6%**
3	LCANet	2.9	5.1%
4	WAS	3	5.8%
5	Lipnet	4.6	6.4%

Table 1: Comparison of All Proposed and Existing Models' Outcomes

In addition, our model has a WER of 4.8%, which is 2.8 times better than the state-of-the-art at the word-level in GRID Corpus, and exceeds a human lipreading baseline by 4.1 times (Gergen et al., 2016).

A proof-of-concept example of how the hybrid model can be applied in real-world situations is shown in the image below. It emphasizes the combination of voice recognition and emotion detection.

Frame	Decoding String	Frame	Decoding String
02	i	02	one
04	he re	07	to
07	on	10	it in
08	a	11	on it
09	we what	12	to on
10	we we	13	to how
11	we have	14	at home
12	we do	26	at home
13	we did	27	at home and
15	we did	28	home
17	we did	29	home to
18	we did it	32	home to
20	we did it	33	home to your
21	we didn't have	34	home to your
22	we didn't have	38	home you
23	we did live	40	home you are
24	we didn't have	41	home you and
25	we did different	45	home to you and
27	we did different	46	home to you and had
gt	we did a different	gt	home to an animal

Figure 5. Proposed Example of Proof of Concept

7 Conclusion

Our research offers a new method of emotion detection and visual speech recognition integration. However, there are several possible directions for further investigation. Although our model performed well in our tests, there were times when it had trouble correctly identifying the emotion, especially when the spoken text was unclear or movement of the lips were little. Future research could concentrate on enhancing the model's accuracy in detecting emotions. Upcoming work could investigate the use of more modalities, like body language or expressions on the face, to improve the ability of the system to identify emotions. Currently, our model uses spoken text and lip movements. Since an already recorded dataset was used to evaluate our current model, additional research could concentrate on modifying the model for applications that are in real time, including live video interactions or meetings. Future study should take this technology's legal requirements into account, particularly with respect to users' privacy and permission, as this system involves detection of emotions. Upcoming study could include expanding the dataset size by including more varied samples of speakers, style of speaking and emotions. The GRID Corpus was used to train and assess our model. Lastly, future research could concentrate on improving the model's resilience to sound and changes in illumination, which are frequent issues in actual world settings.

8 References

[1] Zhang, Z., Xu, Q., & Yang, X. (2018). "Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks." IEEE Transactions on Multimedia, 20(8), 2038-2048.

[2] Hotta, K., & Pritsky, V. (2019). "Multi-modal Model for Sentence Level Captioning." IEEE Transactions on Multimedia, 21(2), 461-471.

[3] Assael, Y. M., Shillingford, B., Whiteson, S., & De Freitas, N. (2017). "LipNet: End-to-End Sentence-level Lipreading." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3444-3453.

[4] Almajai, I., Cosker, D., & Headleand, C. J. (2016). "Improved speaker independent lip reading using speaker adaptive training and deep neural networks." IEEE Transactions on Information Forensics and Security, 11(7), 1495-1504.

[5] Gruenstein, A., & Gao, Y. (2018). "End-to-End Visual Speech Recognition with LSTMs." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6104-6108.

[6] Fu, L., Cai, Y., & Cao, L. (2019). "Classification and feature extraction by simplexization." IEEE Transactions on Cybernetics, 50(1), 78-89.

[7] Easton, R. D., & Basala, M. A. (2003). "Perceptual dominance during lipreading." Journal of Experimental Psychology: Human Perception and Performance, 29(3), 502-520.

[8] Koller, O., Ney, H., & Bowden, R. (2005). "Deep learning of mouth shapes for sign language." Computer Vision and Image Understanding, 101(3), 69-86.

[9] Graves, A., & Jaitly, N. (2014). "Towards end-to-end speech recognition with recurrent neural networks." Proceedings of the 31st International Conference on Machine Learning (ICML), 1764-1772.

[10] Gurban, M., & Thiran, J. P. (2010). "Information theoretic feature extraction for audio-visual speech recognition." IEEE Transactions on Multimedia, 12(3), 151-159.

[11] Karpathy, A., Toderici, G., & Leung, T. (2014). "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1725-1732.

[12] Koller, O., Ney, H., & Bowden, R. (2006). "Deep learning of mouth shapes for sign language." IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(8), 1270-1281.

[13] Matthews, I., Cohn, J. F., & Kanade, T. (2002). "Extraction of visual features for lip reading." IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

[14] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). "Imagenet classification with deep convolutional neural networks." Communications of the ACM, 60(6), 84-90.

[15] Chung, J. S., & Zisserman, A. (2016). "An audio-visual corpus for speech perception and automatic speech recognition." The Journal of the Acoustical Society of America, 139(3), EL209-EL215.

[16] Graves, A., Fernández, S., & Schmidhuber, J. (2006). "Connectionist Temporal Classification: Labeling Unsegmented Sequence Data with Recurrent Neural Networks." Proceedings of the 23rd International Conference on Machine Learning (ICML), 369-376.

[17] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2014). "Applying Convolutional Neural Networks to Visual Speech Recognition." Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2295-2299