Pre - Print Version

Automated Lip Reading to Predict Visemes using Multimodal Convolutional Neural Network with Audio-Visual Features

Khalid Mahboob[1], Umm-e-Laila[2]*, Sana Alam[3],Muhammad Abbas[4], Muhammad Asghar Khan[5], Sidra Fatima[6]

***Abstract*: The process of interpreting sentences based on the movements of a speaker's lips is referred to as lip reading. Traditionally, this task has been approached in two stages using conventional methods: first, by generating or learning audio-visual features, and second, by making predictions. While contemporary deep lip reading techniques benefit from end-to-end trainable datasets, much of the existing research on these models tends to concentrate solely on word classification rather than predicting sequences at the sentence level. Long sentences may be lip-read by humans, as studies have shown. This study emphasizes the value of temporal considerations by highlighting the components that are important for capturing temporal context in instances when communication channels are unclear. In the paper, a lip-reading system for viseme prediction is shown. The system uses a Convolutional Neural Network (CNN) with a recurrent network, spatiotemporal convolutions, and the connectionist temporal classification loss. A variable-length series of video frames is efficiently mapped to text using an end-to-end training procedure. Both visual and auditory qualities are evaluated using the CNN architecture. The CNN model outperforms trained human lip readers and achieves accuracies of 72.8% CER and 80.8% WER (unseen speakers with audio), whereas 46.2% CER and 56.6% WER (unseen speakers without audio), which are reasonable accuracies on the GRID corpus by splitting test at the level of the sentences.**

***Keywords:* lip, reading, model, visemes, accuracy, convolutional neural network**

[1-2-3-4-5]College of Computer Science and Information Systems, Institute of Business Management, Karachi, Pakistan |
[6] Computer Engineering Department, Sir Syed University of Engineering & Technology, Karachi
Country : Pakistan
Email: umme.laila@iobm.edu.pk

INTRODUCTION:

While deep learning techniques, like Convolutional Neural Networks (CNNs), are excellent at deriving meaning from heterogeneous or ambiguous data, they are also very good at identifying complex patterns and trends that are beyond the capacity of human or other computer systems to recognize. A CNN model can be pondered an expert in the particular data group it was trained to observe after training [1].

Without contextual information, deciphering lip movements poses a considerable challenge for humans. Lip reading involves not only observing the lips but also discerning subtle movements of the tongue and, at times, the teeth. Many lip-reading cues remain latent and are difficult to interpret within a given context. CNN-driven Lip Reading, a direct method for speech recognition, stands out as an advanced technology frequently applied in this domain. We introduce a lip-reading approach that takes into account both lip shape and the intensity of the mouth region. Tracking and parameterizing the inner and outer boundaries of the lips in a series of images yield shape information. Based on grey-level data, a principal component analysis model is used to extract intensity details [2].

In contrast to earlier methods, our approach involves the concurrent deformation of both the intensity area and the shape model to guarantee the replication of identical object features following non-rigid deformation of the lips. We present recognition methods for speaker-independent applications utilizing these attributes. Initial findings suggest that combining shape and intensity information may yield enhanced performance, while comparable results can be achieved by employing either one of them individually [3].

1.1. Significance of Speech Recognition:

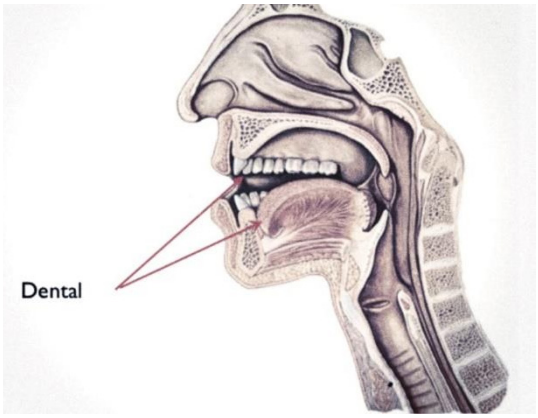
The discipline of speech recognition is identified as the major compelling domains of computer science. It is essential that computers be able to comprehend both speech and gestures. The difficulty arises in recognizing diverse and complex terms. Convolutional Neural Networks (CNNs), designed comprises of the functioning of the human brain, prove effective in various pattern recognition tasks and demonstrate exceptional learning capabilities [4].

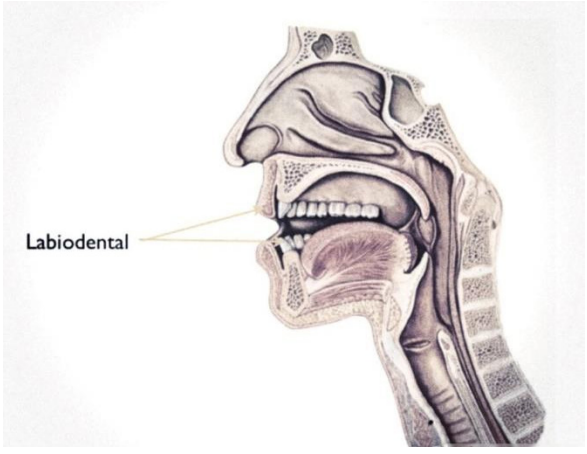
CNNs may combine several heterogeneous input features that don't have to be considered independent to identify the best combination of these characteristics for the task of

classification. This research aims to use CNN's ability to increase speech recognition accuracy at the sentence level using visual features [5].

1.2. Process of Lip-Reading:

Lip reading is the technique of extracting visual traits of speech from a person's lips. Though it has also been shown that information regarding the shape of the tongue and teeth may also transmit substantial speech cues, the curves of the inner and outer lips contain the most essential visual speech signals. Fricatives feature easily distinguishable articulation sites, such as upper teeth on the lower lip, interdentally (tongue behind front teeth), and alveolar (tongue contacting gum ridge). Wrinkles and lip protrusion may provide extra information during speaking [6]. Figure 1 below depicts the human mouth with arrows indicating the English articulation points.





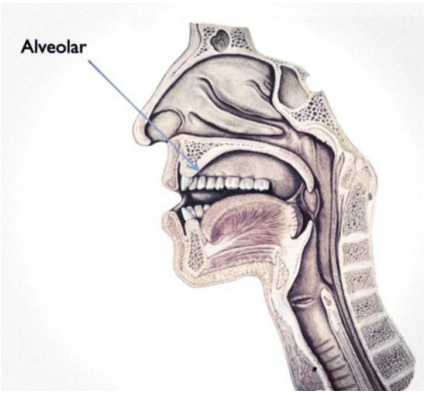


Figure 1. Human mouth representation with articulation points

In English, fricatives (see Table 1) are produced by slightly inhibiting airflow via the mouth. They are distinguished by phonological features: sibilance, point of articulation, and voicing. The articulation points can be either voiced or voiceless. Fricatives are identified by a variety of cues [7].

Table 1. List of fricatives in English language w.r.t point of articulation

Point of Articulation	Voiceless		Voiced	
	IPA	Examples	IPA	Examples
Labiodental	/f/	"fro", "fat"	/v/	"vine", "van"
Interdental / Dental	/θ/	"thick", "think"	/ð/	"the", "this"
Alveolar	/s/	"suit", "sun"	/z/	"zit", "zoom"

Lip reading techniques may be divided into two categories:

1.-based techniques.
2. Model-based techniques.

Grey-level data from an image area including the lips is utilized for speech parameters in image-based systems, either directly or after some processing. This preserves the majority of the visual data; nevertheless, the identification algorithm still needs to separate audio recordings from language and illuminating variability [8].

Lips are frequently represented in model-based systems by geometric measurements such as the height or width of the outer or inner lip limits, or by a parametric contour model that illustrates the lip borders. The characteristics that were recovered are low-dimensional and insensitive to light.

Model-based systems depend on the user describing characteristics associated with speech. Because of this, the definition could not include all speech-related details and characteristics that are challenging to illustrate, including the exposure of the tongue and teeth [9, 10]. Hence, the prime objective of this work is to automate lip reading for the purpose of predicting visemes. This innovation possesses major value for numerous applications such as enhanced hearing aids, biometric verification, security systems, discreet dictation in public areas, speech recognition in loud circumstances, and the processing of silent movies, among others. The remaining portion of this article is divided into various sections as follows: Section 2 provides a review of related works on automatic lip reading, while Section 3 delves into the detailed explanation of the CNN architecture. The processes of data pre-processing and methodology are elucidated in Section 4. Section 5 comprises of presenting and assessing the outcomes along with a comprehensive discussion. Finally, Section 6 encapsulates our concluding remarks.

LITERATURE REVIEW

Many techniques for reconstructing speech from silent videos have recently been investigated by researchers. For machine lip-reading, it is essential to extract spatiotemporal information from movies, which is difficult because both location and motion are important. The goal of modern deep learning techniques is to fully capture these properties. Nevertheless, most previous research only addresses word-level sequence prediction, not sentence-level sequence prediction. As a result, we suggested an automatic lip-reading system that makes use of visual characteristics to scan a user's lip movements and make an educated guess as to the visemes the user is expressing. This section's linked works will assist in identifying any research gaps using current methods.

An automated lipreading system employs a recurrent network of spatiotemporal convolutions. An end-to-end model with the connectionist temporal classification loss has been developed, depending upon deep learning convolutional neural networks. This loss is used to translate a text sequence from a variable-length video frame series. Employing characteristics extracted from films, the efficacy of the trained lip-reading method in sentence prediction was examined. The study discovered that sufficient information about the speaker was captured by the facial landmark representation. Nevertheless, the important aspects of lip-reading that might be present in films are not captured by this interpretation [11].

The principal objective of a study is to form a network architecture for data recognition, processing, and acquisition in lip-reading. A research was conducted using an algorithm for lip reading that was both accurate and adaptive. To extract and segment the mouth region, a planned hybrid model with a newly proposed edge centered on a proposed filter was first implemented. Next, convolutional neural networks (CNN) and bi-directional gated recurrent units (Bi-GRU) were used to train the spatiotemporal model. In testing algorithms, an accuracy of 90.38% has been achieved altogether. The system's performance demonstrated lip segments by applying lip segmentation as input to the proposed Spatiotemporal model [12].

This report extensively examines temporal models and data augmentation techniques applied to the LRW dataset. It illustrates how the synergistic application of optimal augmentations and training approaches can yield cutting-edge performance. The study reveals that among various augmentations, time masking emerges as the most crucial, followed by mixup. Additionally, the study identifies Densely-Connected Temporal Convolutional Networks (DC-TCN) as the most efficient temporal model for lip-reading isolated words. Integration of these methods results in an impressive 93.4% classification accuracy, surpassing the current state-of-the-art LRW dataset results by 4.6%. Further enhancements to a 94.1% accuracy can be achieved through pre-training on new datasets. An error analysis of diverse training methods highlights performance improvements, particularly in accurately classifying challenging words [13].

Lip reading involves converting video data into textual information. The proposed technique comprises a test dataset, image frame analysis, and the generation of text output based on detected words. The test dataset was constructed by integrating various facial expressions associated with different words. The system consists of four key components: test data, data preprocessing, word identification, and output generation. In the test data component, raw video clips serve as input, which are then divided into frames and preprocessed during the data preprocessing step. These processed frames are compared with training data during the word identification stage. Lip-Interact illustrates how the camera functions as a sensor, identifying lip gestures by capturing every movement in the video. Notably, there is a time gap between the capture and utilization of each image. It's important to note that this technique may encounter challenges and potential failures for various reasons. The lip gesture for comparable phrases is one of the most typical issues researchers have encountered. It might be difficult to tell what the speaker is attempting to say vs what the system truly recognizes. However, employing machine learning makes it feasible to decrease the error associated with this problem with sufficient training and sample data. The accuracy achieved in this study was found to be 90% [14].

In visual speech recognition (VSR), speech is synthesized by analyzing the motions of the tongue and teeth using visual input. Recently, deep learning has shown to do exceptionally well in VSR, surpassing lip-readers on benchmark datasets. Nonetheless, a few problems persist with VSR systems. The

distinction between homophones, or words that sound similar, is a significant issue that contributes to word ambiguity. Words like "a," "an," "eight," and "bin" require more visual information to be learned than words with lengths of less than 0.02 seconds, which is another technical flaw in standard VSR systems. This study presents a novel lip-reading architecture consisting of three independent convolutional neural networks (CNNs): a multi-layer feature fusion CNN, a three-dimensional CNN, and a densely linked CNN. A two-layer bi-directional gated recurrent unit is placed after the three CNNs. To train the whole network, connectionist temporal classification was employed. For the unseen-speaker dataset, the suggested architecture reduced the baseline model's character and word mistake rates by 5.681% and 11.282%), respectively, based on accepted automatic speech recognition assessment standards. However, automated speech recognition using just VSR remains difficult since speech combines both visual and aural information [15].

Densely Connected Temporal Convolutional Network (DC-TCN) is used for isolated word lip-reading. Temporal Convolutional Networks (TCN) have shown excellent results recently in a range of vision tasks; nevertheless, they still lack the receptive fields required to capture the complex temporal dynamics observed in lip-reading scenarios. Dense connections are added to the network to overcome this obstacle and better capture temporal information. Furthermore, the Squeeze-and-Excitation block, a basic attention strategy, has improved the model's classification performance. The DC-TCN methodology achieved accuracies of 88.36% on the Lip Reading in the Wild (LRW) dataset and 43.65% on the LRW-1000 dataset, respectively, outperforming all traditional methodologies, according to the study's findings [16].

The goal of the study is to help those who are hard of hearing comprehend interaction without the need for specialized instruction or assistance from others. The study concentrated on speech recognition difficulties such as homophones and co-articulation. The problem can be solved with the help of deep learning with long-short-term memory, and the procedure can be enhanced by combining it with facial feature extraction. Depth-sensing and color imaging work together to further increase the classifier's accuracy. The core of the system was built in Python using the TensorFlow and Keras libraries. The photos were processed using OpenCV. 3000 examples from the MIRACL-VC1 dataset have been utilized. According to the findings, the LSTM method produced results with an accuracy of 85% [17].

To simulate real environments, a two-stage corrector using Generative Adversarial Networks and a lip deflection classifier using Convolutional Neural Networks were created. A challenging dataset with many speakers to simulate a dynamic environment and substantial lip deflection angles was presented. These models are used to correct lip deflection and improve recognition accuracy. The proposed network is successful in tackling significant lip deflection angles in real-world settings, as evidenced by absolute improvements of 18.3% and 7.4% in compared to scenarios without preprocessing and confined to face alignment alone [18].

A lip-reading system based on neural networks has been developed to predict sentences from silent conversations in movies, encompassing a diverse vocabulary. This system exhibits resilience to fluctuations in lighting conditions, operates without reliance on a specific lexicon, and relies solely on visual data represented by visemes, capturing a limited set of distinct lip movements. The performance criterion of the model has been validated using the crucial BBC Lip Reading Sentences 2 (LRS2) benchmark dataset, demonstrating a notable 15% enhancement compared to its previous performance. Experimental findings indicate that the proposed model maintains its effectiveness even when exposed to variations in illumination during movie sequences. Notably, the classification accuracy for visemes in this approach surpassed 95%, but the conversion process resulted in a substantial reduction in the classification accuracy of words, as revealed by the results [19].

The issue of automatic face recognition for individuals engaged in speaking activities was addressed in this study. The investigation aimed to determine whether incorporating additional signals related to articulation, alongside facial features, could enhance the precision of emotion identification when integrated into a deep neural network (DNN) model. To discern the facial expressions of speaking participants, a spatiotemporal Convolutional Neural Network (CNN) and a Gated Recurrent Unit (GRU) cell Recurrent Neural Network (RNN) were developed utilizing the RAVDESS dataset. Initially, these models were trained solely on facial features, and subsequently on signals related to articulation extracted from a lip-reading model, as well as on a combination of facial features and varying numbers of consecutive frames included in the input. The study findings reveal that integrating articulation features into DNNs enhances classification accuracy by up to 12%, with a more pronounced improvement observed when more consecutive frames are incorporated into the model's input [20].

CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE:

Convolutional neural networks, sometimes described as CNNs or ConvNets, are a particular kind of neural networks that are principally good at processing input that has a grid-like structure, like an image. A digital picture is a grid-like arrangement of pixels that functions as a binary illustration of visual data. As seen in Figure 2 [15], each pixel is given a pixel value that represents its brightness as well as color.

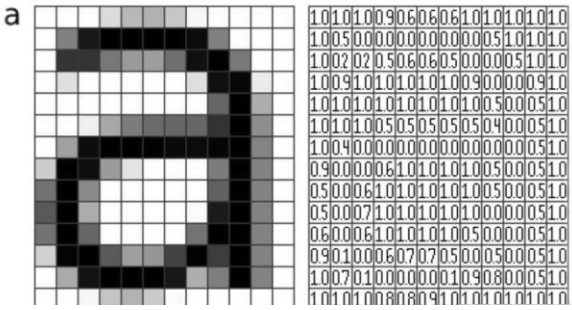


Figure 2. Image representation using a grid of pixels

When the human brain comes into contact with a picture, it begins to investigate a large amount of data. Each neuron in this complex neural network has a distinct receptive field that links to other neurons to form a total receptive field that encompasses the whole visual field. Similar to the human visual system, each neuron in a Convolutional Neural Network (CNN) only processes data inside its assigned receptive area. The network's first layers identify basic components like lines and curves before moving on to more intricate aspects like faces and objects. Computers can be given the capacity to see by using a CNN [1].

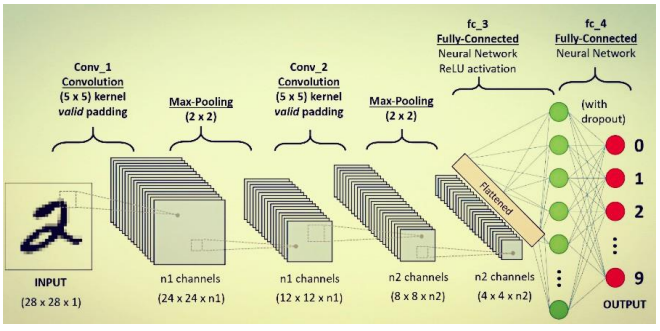


Figure 3. CNN architecture representation

3.1. Convolution Layer

The cornerstone of CNNs is the convolution layer, which shoulders the basic computational load within the network. This layer performs a dot product operation between two matrices: one is the kernel, containing adjustable parameters, and the other represents the limited region of the receptive field. Although the kernel is more detailed than an image, it is spatially smaller. To clarify, in the context of an image with three (RGB) channels, the kernel's height and width are spatially compact, while its depth extends to encompass all three channels [21].

The convolution layer, which bears the majority of the network's computational load, is the foundation of CNNs. This layer carries out a dot product operation between two matrices, one of which represents the restricted area of the receptive field and the other is the kernel with modifiable parameters. The kernel is smaller in space than an image but has greater detail. To be more precise, the kernel's depth spans all three channels in the context of an image with three (RGB) channels, although its height and breadth are spatially compact [21].

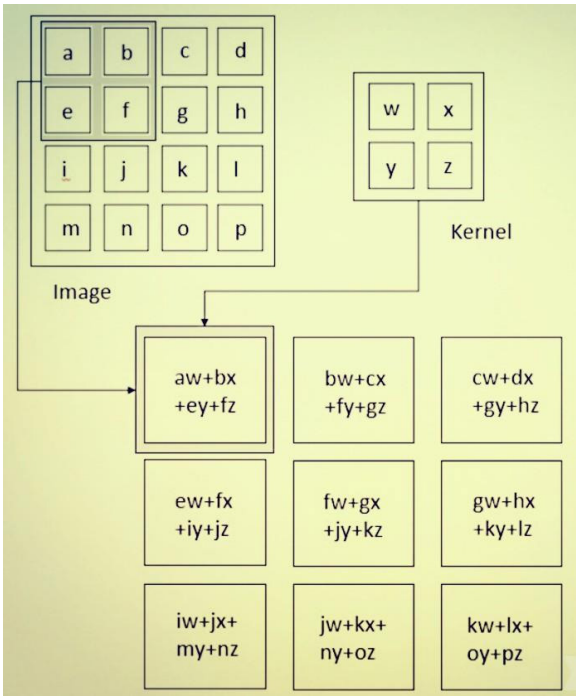


Figure 4. Convolution process representation

3.2. Pooling Layer

The pooling layer substitutes the network's output at specific periods by generating a summary statistic from the surrounding outputs. This reduces the spatial dimension of the representation, hence reducing the computational load and required weights. The pooling technique is applied independently to each model slice [23].

Pooling processes include a weighted average based on the distance to the center pixel, the rectangular neighborhood standard, and the rectangular neighborhood L2 norm. However, the most common method is max pooling, which finds the neighborhood's largest value. The following formula may be used to calculate the output volume's dimensions if there is an activation map with dimensions W x W x D, a spatially sized pooling kernel F, and a stride S.

$$out = \frac{W-F}{S} + 1 \quad \text{.....(1)}$$

As a result, the output volume will be Wout x Wout x D. Because of certain translation invariance, pooling ensures that an item is always discernible no matter where it is in the frames (see Figure 5) [24].

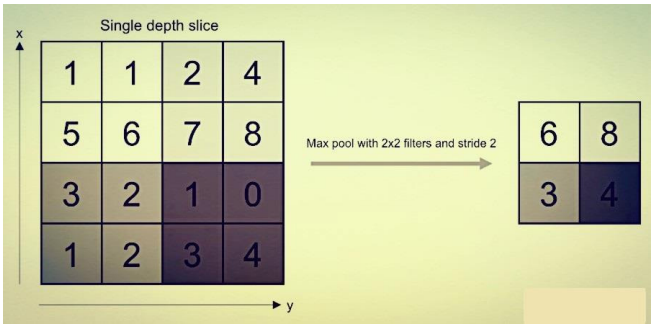


Figure 5. Pooling process representation

3.3. Fully Connected Layer

Every neuron in a layer of a conventional Fully Connected Neural Network (FCNN) is closely connected to every other neuron in the layer above and below it. This type of connection enables calculation using the standard method of matrix multiplication together with bias inclusion. As shown in Figure 6 [25], the Fully Connected (FC) layer is essential for enabling the mapping of representations between the input and output.

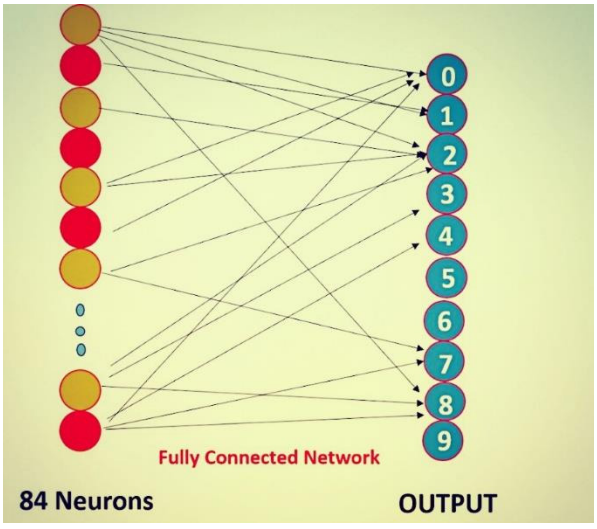


Figure 6. Predicted number by output neurons using softmax.

DATA PRE-PROCESSING AND METHODOLOGY:

With just the speaker's lip movements and visemes known, as well as the possibility of a mistake or misclassification in the input viseme sequence, the aim of this study is to predict an expected phrase that a subject is likely to say given a series of visemes. Many lip-reading datasets exist, however, the majority either contain one word or are far too short of sentences [32]. The GRID corpus comprises 34 speakers' audio and video records, each of whom produced 1000 utterances and 34000 visemes. While whole sentences are included in the GRID corpus, we will only address the simpler case of single-word prediction. This work uses temporal context to predict sequences, which improves accuracy [27].

The GRID corpus is used here to evaluate this study as it is sentence-level and has the most data. Due to the absence of speaker 21's movies and the corruption or futility of a few others, there are just 32746 working videos. Three male and three female speakers' data are awaited for analysis, with 70% of the data being used for training with 22922 films and 30% being used for testing with 9824 movies (unseen speakers with and without audio) that haven't been used in the literature before [27]. The dataset has been split into separate video clips for the pronunciation of each number, with American English serving as the baseline. 68 landmarks were used in the analysis of the movies using the DLib face detector, the online Kalman filter, and the iBugfacelandmark predictor (refer to Figure 7). Thanks to these landmarks, we can extract a mouth-centered crop with a size of 100×50 pixels every frame using an affine transformation. We equalize the RGB channels throughout the training set in order to attain zero and one unit variance [15, 28].

Using a DLib face detector, the targeted face and mouth are identified at the pre-processing data stage. When creating the bounding box for the lips, the algorithm generates the (x, y) coordinates of the diagonal edges. To mitigate overfitting, we introduce essential modifications to the dataset. Initially, standard and horizontally mirrored image sequences are employed for the initial training. Consistent pre-processing and augmentation procedures, as outlined in the GRID dataset [11], are applied to train and test all models.

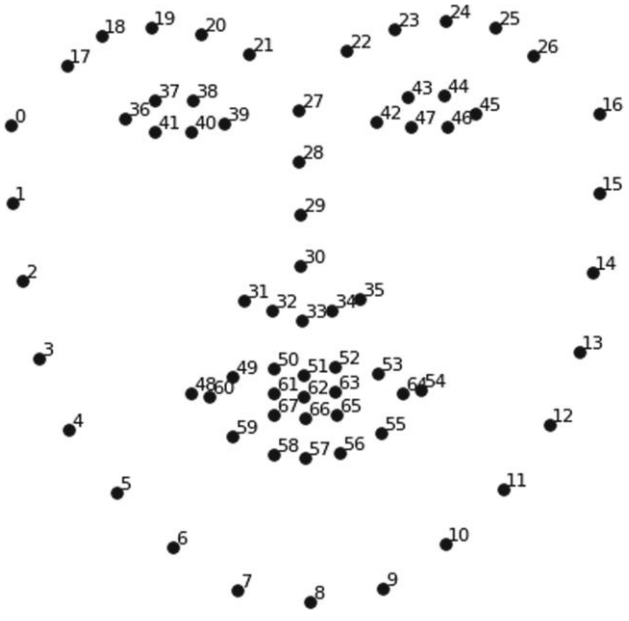


Figure 7. 68 Landmarks representation identified with DLib's face detector

The CNN approach predicts the speaker's vocabulary based on a video input. The model will forecast data from the grid dataset since it was trained on it. The model consists of two

Bi-GRUs layers, three spatiotemporal convolution layers, channel-wise dropout layers, spatial max-pooling layers, and softmax activation functions and rectified linear unit (ReLU) activation functions for sequence classification. In order to evaluate the character error rate (CER) and word error rate (WER) scores using CTC beam search, all models were built with Keras with a TensorFlow backend and TensorFlow-CTC decoder [33]. The Rectified Linear Unit (ReLU) performs the computation of the function f (k) = max (0, k). In other words, there is a zero threshold at which the activation is present. The following is how the CNN steps are used:

→ [CONV 1] → [BATCH NORM] → [ReLU] → [POOL 1]
→ [CONV 2] → [BATCH NORM] → [ReLU] → [POOL 2]
→ [FC LAYER] → [OUTPUT]

One stride, two padding, and a 5 x 5 spatial kernel are employed for each of the two convolutional layers. The max pool operation is performed with a kernel size of 2, stride 2, and zero padding for both pooling levels. ZeroPadding3D layer is utilized to pad zeros in three dimensions. The Batch Normalization layer equalizes the input and prevents improper weight matrix initialization by explicitly compelling the network to adopt a Gaussian unit distribution [33, 34]. The main procedure for lip-reading using CNN in this work is depicted in Figure 8.

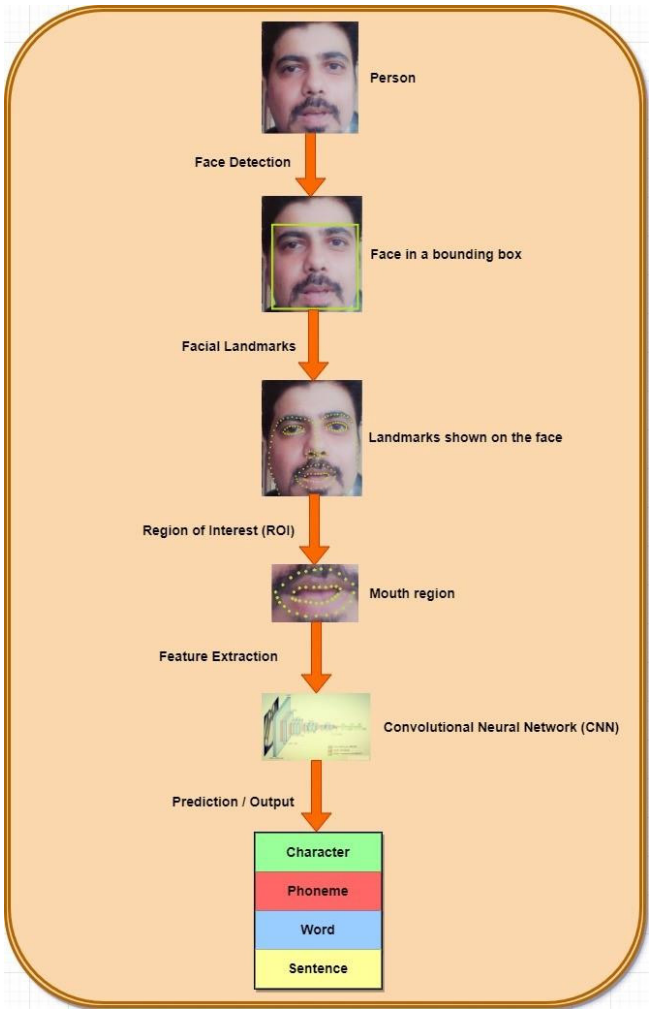


Figure 8. The representational flow of lip-reading process

RESULTS AND DISCUSSION:

The dataset employed was the GRID corpus. Included are videos of the 34 speakers—both male and female—uttering 1,000 phrases. The grammatical structure uses the following: command (4) + color (4) + preposition (4) + letter (25), digit (0) + adverb (4). The number denotes the number of word options available in each of the six-word categories. 64000 possible phrases may be generated by combining the following categories: {bin, lay, place, set}, {blue, green, red, white}, {at, by, in, with}, {A,...,Z}\{W}, {0,…,9}, and {again, now, please, soon}. We compute the word error rate (WER) and character error

deletions, insertions, and substitutions necessary to convert the amount of words (or characters) in the ground truth divided by the forecast into the ground truth. When the predicted and actual sentences have the same number of words as the ground truth, WER is typically equal to classification error. This is especially true in our situation because practically all errors are substitution errors. Equations (3) and (4) are used to calculate the CER and WER, where C and W stand for characters and words, respectively [35].

$$(\%) = (\frac{C+C_D+C_I}{C}) \times 100 \quad \text{.....(3)}$$
$$N$$

$$(\%) = (\frac{W+W_D+W_I}{W}) \times 100 \quad \text{.....(4)}$$
$$N$$

The entire edit distance was computed in order to obtain the error rate measurements for accuracy evaluation and convert them into percentages. The formula, where D is the number of deletions that should have been made from the decoded face characters and I is the number of characters, is provided [28]. Characters were utilized in place of the inaccurate classifications, and N is the total number of characters in the ground truth. S is the number of characters in the ground truth.

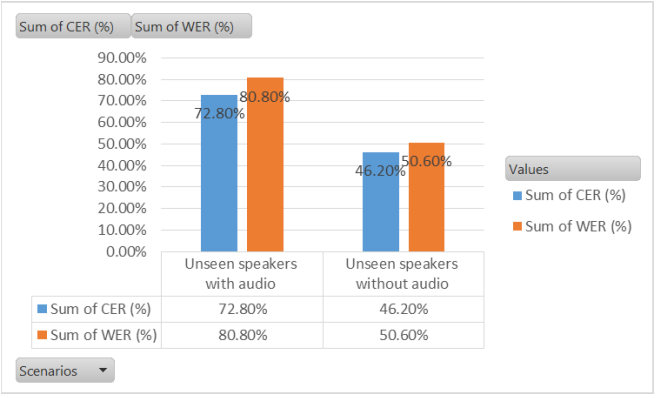


Figure 9. Comparative representation of Accuracies w.r.t CER and WER

As anticipated, the fixed sentence arrangement and the small selection of words at each place in the GRID corpus make it easier to employ context, which improves performance. On the unseen speakers without an audio split, the accuracy achieved is 46.2% CER and 50.6% WER, respectively, and the accuracy achieved with the unseen speakers with an audio split is 72.8%CER and 80.8% WER, respectively (see Figure 9). It is important to note that the maximum accuracy is achieved with the unseen speakers with audio compared to the unseen speakers without audio enhancing with the convolutional stack [30].

In this part, we conduct a phonological analysis of our system-learned representations. We

first construct saliency visualizations to present where the system has learned to attend. To produce a CTC alignment, we specifically supply a model an input and voraciously decode a sequence of outputs. Next, we compute the gradient concerning the input video frame sequence using guided backpropagation. Second, utilizing intra-viseme and inter-viseme confusion matrices, we train our system to predict ARPAbet phonemes rather than characters in order to analyze visual phoneme similarity [26].

We interpret the learned behavior using typical visemevisualization approaches
demonstrating that the model pays attention to phonologically significant places in the video. We refer to the lip images retrieved for each phoneme in our work as visemes (see Figure 10). Secondly, we build a morph transition from each viseme image to each other viseme image in a manner that is depicted in a figure. With the support of this transformation, we can create images of intermediate lip shapes between any two visemes, allowing for a smooth and accurate transition between them. We define such transformations for N visemes in our final viseme collection i.e. N2. Finally, we concatenate viseme morphs to create a unique visual utterance [31].

,

/p, b, m, v/

/o, w/

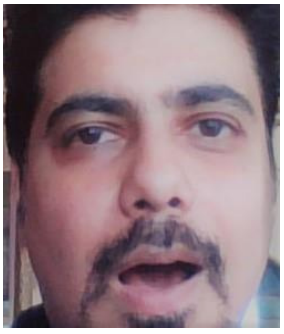
/a, r/

/q, u/

/d, j, k/

/c, f, g, l, s, z/

/e, h, n, t, x/



/i, y/

Figure 10. Video frames representation of mapping phonemes-to-visemes

Much articulatory movement is needed to produce the first part of the word please: the lips must be tightly pulled together to create the bilabial plosive /p, b, m, v/. (frame 1). In preparation for the upcoming lateral /c, f, g, l, s, z/, the tongue's blade touches the alveolar ridge simultaneously. Then, the lips separate, enabling the trapped air to escape (frame 6). For the close vowel /i, y/ (frame 8), the jaw and lips expand even more, as seen by the increased space between the corners of the mouth and the upper and lower lips' midpoints. For the remainder of its duration (frames 2-4), during which the attention level significantly decreases, this vowel's lip position remains stable since it is a reasonably steady-state vowel. When the tongue blade has to be near to the alveolar ridge for the sounds mentioned in a figure (frames 5 and 7), the jaw and lips are then slightly open [29].

We employ phoneme-to-viseme mapping for our study, grouping the phonemes into

the following categories: Alveolar-semivowels (A), Alveolar-fricatives (B), Alveolar (C), Palato-alveolar (D), Bilabial (E), Dental (F), Labio-dental (G), Velar (H), and Lip-rounding based vowels (V). The GRID corpus includes 31 of the ARPAbet's 39 phonemes. In order to organize phonemes into viseme clusters, confusion matrices between phonemes are first computed (see Figures 11— 15).

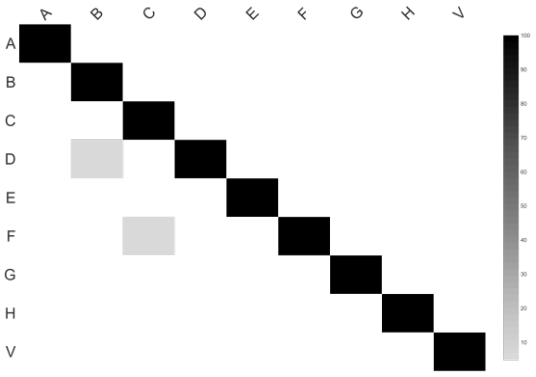


Figure 11. Confusion matrix of viseme categories

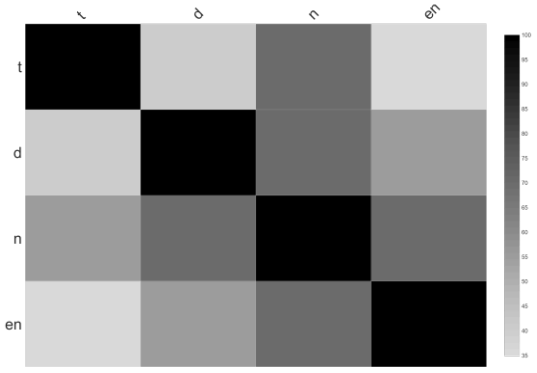


Figure 12. Intra-viseme and inter-viseme confusion matrix of Alveolar

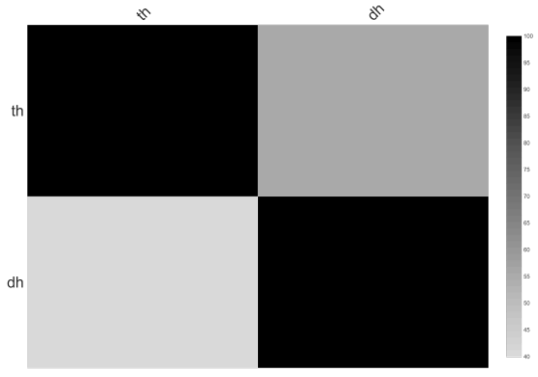


Figure 13. Intra-viseme and inter-viseme confusion matrix of Dental

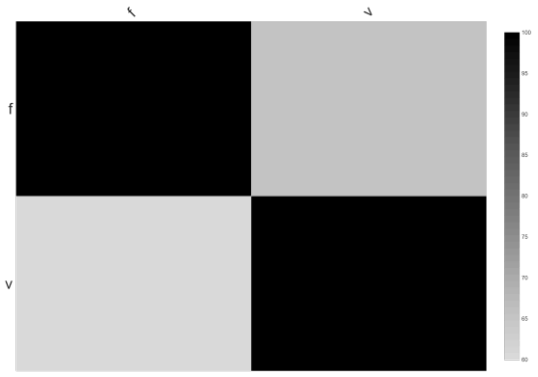


Figure 14. Intra-viseme and inter-viseme confusion matrix of Labiodental

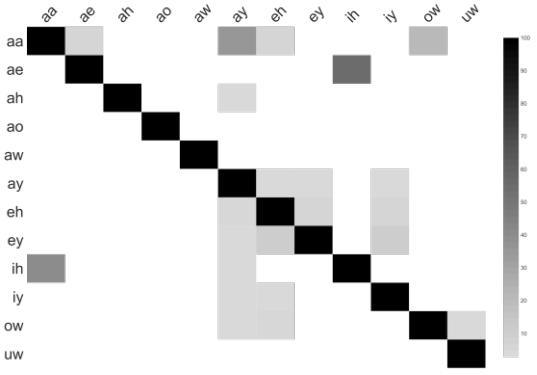


Figure 15. Confusion matrix of Lip-rounding vowels

The Figure 11's diagonal confusion matrix and the extremely minor misinterpretation between the palato-alveolar (D) and alveolar (C) visemes attest to the correctness of the s viseme classification. The sole articulatory difference between the palato-alveolar /sh zh/ and alveolar /s z/ fricatives is a little shift in tongue position, which is contrary to the palate just beyond the alveolar ridge, which is hard to notice from the face. Alveolar /t/ and dental /th/ both suit this pattern.

It is not unexpected that the categories of alveolar stops (/t d n en/), dental stops (/th dh/), and labiodental stops (/f v/) (see Figures 12-14) are confused because they seem to be almost similar when completely closed at the same point of articulation. From the front, it is impossible to distinguish between vocal fold vibration and velum motion.

The reason /aa/ and /ay/ are often mistaken is because the first element and most of the fricative /ay/ are articulatory identical to /aa/—an open back unrounded vowel (see Figure 15). The uncertainty between the relatively close vowel /ih/ and the very open vowel /ae/ may seem unexpected at first, but in the sample, /ae/ only occurs in the word at a function word that is generally pronounced with a decreased, weak vowel /ah/. The most common unstressed vowels are /ah/ and /ih/, and there is significant variation in both [31].

CONCLUSION

We presented a Convolutional Neural Network (CNN) based automatic lip reading system, which is the first deep learning application for comprehensive learning of a model that transforms image frame sequences storing a speaker's mouth into whole sentences. By using the end-to-end approach, predicting sentences does not need splitting out movies into individual words. Our empirical analysis highlights the significance of effective temporal aggregation and spatiotemporal feature extraction. Furthermore, our system achieves 72.8% CER and 80.8% WER (unseen speakers with audio) accuracy, which is much better than the baseline performance of a human lip reader. In comparison, the word-level advanced in the GRID corpus only manages 46.2% CER and 56.6% WER. As opposed to unseen speakers without audio enhancement using the convolutional stack, it is crucial to remember that the highest accuracy is obtained with unseen speakers with audio.

Although the proposed approach is empirically successful, extensive research on voice recognition indicates that more data can improve performance even further. In the following work, we want to substantiate this theory using methods other than CNN on large-scale datasets at the sentence-level level. In some situations, using video alone is imperative, especially when silent notation is involved.

Declaration of Competing Interest:

The authors declare no competing interests associated with this work.

REFERENCES:

[1] Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and LearningSystems. 2021; 1–21. Doi: https://doi.org/10.1109/tnnls.2021.3084827.

[2] Pujari S, Sneha S, Vinusha R, Bhuvaneshwari P, Rashaswini C. A Survey on Deep Learning based Lip-Reading Techniques. Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). 2021; (Icicv):1286– 93.

[3] Oghbaie M, Sabaghi A, Hashemifard K, Akbari M. Advances and Challenges in Deep Lip Reading. Computer Vision and Image Understanding [Internet]. 2021; 1–31. Available from: http://arxiv.org/abs/2110.07879.

[4] Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence. 2020; 9(2):85–112. Doi: https://doi.org/10.1007/s13748-019-00203-0.

[5] Fenghour S, Chen D, Guo K, Li B, Xiao P. Deep Learning-Based Automated Lip-Reading: A Survey. IEEE Access. 2021; 9:121184–205. Doi: https://doi.org/10.1109/ACCESS.2021.3107946

[6] Nandini M.S., Bhajantri N.U. A comprehension of contemporary effort for tracking of lip.

International Journal of Bioinformatics Research and Applications. 2020; 16(1):85. Doi: https//doi.org/10.1504/IJBRA.2020.104857.

[7] McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychol Rev. 2011 Apr; 118(2):219-46. Doi: https//doi.org/10.1037/a0022325.

[8] Fenghour S, Chen D, Guo K, Li B, Xiao P. An effective conversion of visemes to words for high-performance automatic lipreading. Sensors. 2021; 21(23). Doi: https://doi.org/ 10.3390/s21237890.

[9] Lu Y, Li H. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Applied Sciences (Switzerland). 2019; 9(8). Doi: https://doi.org/ 10.3390/app9081599.

[10] Ivanko D, Ryumin D. Development of visual and audio speech recognition systems using

deep neural networks. CEUR Workshop Proceedings. 2021; 3027:905–16. Doi: https://doi.org/ 10.20948/graphicon-2021-3027-905-916.

[11] Mahboob K, Nizami H, Ali F, Alvi F. Sentences Prediction Based on Automatic Lip-Reading

Detection with Deep Learning Convolutional Neural Networks Using Video-Based Features. Vol. 1489 CCIS, Communications in Computer and Information Science. Springer Singapore; 2021. 42–53 p. Doi: http://dx.doi.org/10.1007/978-981-16-7334-4_4.

[12] Miled, M., Messaoud, M.A.B. & Bouzid, A. Lip reading of words with lip segmentation and deep learning. Multimedia Tools and Applications (2022). Doi: https://doi.org/10.1007/s11042-022-13321-0.

[13] Ma P, Wang Y, Petridis S, Shen J, Pantic M, Ai M. Training Strategies for Improved Lip-Reading. ICASSP

2022-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022; 8472–6. Doi: https://doi.org/10.1109/ICASSP43922.2022.9746706.

[14] Chowdhury SMMH, Rahman M, Oyshi MT, Hasan MA. Text Extraction through Video Lip

Reading Using Deep Learning. Proceedings of the 2019 8th International Conference on System Modelingand Advancement in Research Trends, SMART 2019. 2020; 240–3. Doi: https://doi.org/ 10.1109/SMART46866.2019.9117224.

[15] Jeon S, Elsharkawy A, Kim MS. Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors. 2022; 22(1). Doi: https://doi.org/ 10.3390/s22010072.

[16] Ma P, Wang Y, Shen J, Petridis S, Pantic M. Lip-reading with densely connected temporal convolutional networks. Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021. 2021; 2856–65. Doi: https://doi.org/ 10.1109/WACV48630.2021.00290.

[17] Nambeesan AS, Payyappilly C, Edwin JC, P JJ, Alex S. LIP Reading Using Facial Feature Extraction and Deep Learning. International Journal of Innovative Science and Research Technology. 2021; 6(7):92–6.

[18] Zhang C, Zhang S. Lip Reading using CNN Lip Deflection Classifier and GAN Two-Stage

Lip Corrector. Journal of Physics: Conference Series. 2021; 1883(1). Doi: https://doi.org/ 10.1088/1742-6596/1883/1/012134.

[19] Fenghour S, Chen D, Guo K, Xiao P. Lip Reading Sentences Using Deep Learning with only

Visual Cues. IEEE Access. 2020; 8(December):215516–30. Doi: https://doi.org/ 10.1109/ACCESS.2020.3040906.

[20] Bursic S, Boccignone G, Ferrara A, D'Amelio A, Lanzarotti R. Improving the accuracy of

automatic facial expression recognition in speaking subjects with deep learning. Applied Sciences (Switzerland) 2020; 10(11):1–15. Doi: https://doi.org/ 10.3390/app10114002.

[21] Shi YM. Construction of the Convolutional neural network Based on the increase optimizers

provided and atomic layers provided. 2022 2nd International Conference on Consumer Electronics and Computer Engineering, ICCECE2022. 2022; 676–9. Doi: https:///doi.org/ 10.1109/ICCECE54139.2022.9712718.

[22] Jameel SM, Hashmani MA, Rehman M, Budiman A. Adaptive CNN Ensemble for Complex

Multispectral Image Analysis. Complexity. 2020; 2020. Doi: https://doi.org/ 10.1155/2020/8361989.

[23]Wahid A, Umar Khan A, Mukhtarullah, Khan S, Shah J. A MultilayeredConvolutional Sparse

Coding Framework for Modeling of Pooling Operation of Convolution Neural Networks. 2019 IEEE 6th International Conference on Smart Instrumentation, Measurement and Application, ICSIMA 2019. 2019;(August):27–9. Doi: https://doi.org/ 10.1109/ICSIMA47653.2019.9057334.

[24] Jameel SM, Hashmani MA, Alhussain H, Rehman M, Budiman A. An optimized deep convolutional neural network architecture for concept drifted image classification. Vol. 1037, Advances in Intelligent Systems and Computing. 2020. 932–942 p. Doi: https://doi.org/ GRID10.1007/978-3-030-29516-5_70.

[25] Zhu Q, Zu X. Fully Convolutional Neural Network Structure and Its Loss Function for Image Classification. IEEE Access. 2022; 10:35541–9. Doi: https://doi.org/ 10.1109/ACCESS.2022.3163849.

[26] Ting J, Song C, Huang H, Tian T. A Comprehensive Dataset for Machine-Learning-based

Lip-Reading Algorithm. ProcediaComputer Science. 2021; 199:1444–9. Doi: https://doi.org/10.1016/j.procs.2022.01.183.

[27] Huang H, Song C, Ting J, Tian T, Hong C, Di Z, et al. A Novel Machine Lip Reading Model.

Procedia Computer Science. 2021; 199:1432–7. Doi: https://doi.org/10.1016/j.procs.2022.01.18.

[28] Johnston B, Chazal P de. A review of image-based automatic facial landmark identification techniques. Vol. 2018, Eurasip Journal on Image and Video Processing. 2018. Doi: https://doi.org/ 10.1186/s13640-018-0324-4.

[29] Lalitha SD, Thyagharajan KK. A study on lip localization techniques used for lip reading from a video. International Journal of Applied Engineering Research. 2016; 11(1):611–5.

[30] Sukritha SN, Mohan M. Analysis on Lip Reading Techniques and Image Concatenation Concepts. ICCISc 2021 - 2021 International Conference on Communication, Control and Information Sciences, Proceedings. 2021; Doi: https://doi.org/ 10.1109/ICCISc52257.2021.9484920.

[31] Chen T, Rao RR. Audio-Visual Integration in Multimodal Communication. Proceedings of

the IEEE. 1998; 86(5):837–52. Doi: https://doi.org/ 10.1109/5.664274.

[32] Neeraja K, Srinivas Rao K, Praneeth G. Deep Learning based Lip Movement Technique for Mute. Proceedings of the 6th International Conference on Communication and Electronics Systems, ICCES 2021. 2021; 1446–50. Doi: https://doi.org/ 10.1109/ICCES51350.2021.9489122.eff

[33] Jeon S, Kim MS. End-to-End Lip-Reading Open Cloud-Based Speech Architecture. Sensors.

2022; 22(8). Doi: https://doi.org/ 10.3390/s22082938.

[34] Joshi VS, Raj ED. FYEO : A Character Level Model for Lip Reading. 2021 8th International

Conference on Smart Computing and Communications: Artificial Intelligence, AI Driven
Applications for a Smart World, ICSCC 2021. 2021; 257–62. Doi: https://doi.org/ 10.1109/ICSCC51209.2021.9528104.

[35] Sarhan AM, Elshennawy NM, Ibrahim DM. HLR-Net: A hybrid lip-reading model based on deep convolutional neural networks. Computers, Materials and Continua. 2021; 68(2):1531–49. Doi: https://doi.org/ 10.32604/cmc.2021.016509.