ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/369296415

Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network

Article *in* Acoustics · March 2023

DOI: 10.3390/acoustics5010020

CITATIONS
12

READS
128

2 authors:

Shashidhar R
JSS Science and Technology University,Sri Jayachamarajendra College of Enginee…
87 PUBLICATIONS **590** CITATIONS
SEE PROFILE

Sudarshan Patilkulkarni
Sri Jayachamarajendra College of Engineering
42 PUBLICATIONS **366** CITATIONS
SEE PROFILE

All content following this page was uploaded by Shashidhar R on 16 March 2023.

The user has requested enhancement of the downloaded file.

acoustics

MDPI

Article

Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network

Shashidhar Rudregowda [1], Sudarshan Patil Kulkarni [1], Gururaj H L [2,*], Vinayakumar Ravi [3,*] and Moez Krichen [4,5]

[1] Department of Electronics and Communication Engineering, JSS Science and Technology University, Karnataka 570006, India
[2] Department of Information Technology, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal 576104, India
[3] Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar 34754, Saudi Arabia
[4] Department of Information Technology, Faculty of Computer Science and Information Technology (FCSIT), Al-Baha University, Alaqiq 65779-7738, Saudi Arabia or moez.krichen@redcad.org
[5] ReDCAD Laboratory, University of Sfax, Sfax 3038, Tunisia
* Correspondence: gururaj.hl@manipal.edu (G.H.L.); vravi@pmu.edu.sa (V.R.)

Abstract: Visual speech recognition (VSR) is a method of reading speech by noticing the lip actions of the narrators. Visual speech significantly depends on the visual features derived from the image sequences. Visual speech recognition is a stimulating process that poses various challenging tasks to human machine-based procedures. VSR methods clarify the tasks by using machine learning. Visual speech helps people who are hearing impaired, laryngeal patients, and are in a noisy environment. In this research, authors developed our dataset for the Kannada Language. The dataset contained five words, which are Avanu, Bagge, Bari, Guruthu, Helida, and these words are randomly chosen. The average duration of each video is 1 s to 1.2 s. The machine learning method is used for feature extraction and classification. Here, authors applied VGG16 Convolution Neural Network for our custom dataset, and relu activation function is used to get an accuracy of 91.90% and the recommended system confirms the effectiveness of the system. The proposed output is compared with HCNN, ResNet-LSTM, Bi-LSTM, and GLCM-ANN, and evidenced the effectiveness of the recommended system.

Keywords: VGG16; CNN; Kannada; visual speech; custom dataset

check for updates

Citation: Rudregowda, S.; Patil Kulkarni, S.; H L, G.; Ravi, V.; Krichen, M. Visual Speech Recognition for Kannada Language Using VGG16 Convolutional Neural Network. *Acoustics* **2023**, *5*, 343–353. https://doi.org/10.3390/acoustics5010020

Academic Editor: Jian Kang

Received: 12 December 2022
Revised: 21 February 2023
Accepted: 6 March 2023
Published: 16 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Observing the lip movement to understand the pronounced word is called lip-reading. Lip reading performs a major part in understanding speech and communication with humans, especially for hearing impaired people. Lip reading techniques are used in a wide range of contexts, including biometric identification, silent notation, a forensic analysis of surveillance camera recordings, and communication with autonomous vehicles. For a multi-ethnic and bilingual country, such as India, the development and readiness of spoken language corpora in local languages are of the utmost importance. The problems of local bias, inflection, single style, and a variety of other problems connected with every environmental area and language are important influences on the presentation of systems that recognize speech. Lip reading has become a famous area in speech research. Research on lip-reading for identifying human action is vast. This is due to the large amount of phonemes in human linguistic that are visually denoted by a lesser number of visemes. Thus, a similar viseme can be used to signify numerous phonemes, which complicates several lip readers.

Our research aims to predict spoken words from a video of someone speaking without audio and vice versa. The goal of the research is to create an algorithm for visual speech recognition, and the suggested method helps those who have hearing loss.

Just conveying one's thoughts to another person verbally is communication. By observing lip movement, one can visualise spoken words through lip reading. For instance, hearing-impaired people frequently utilize lip reading in daily interactions to comprehend one another in noisy settings and in circumstances where the audio speech signal is difficult to grasp. Hearing-impaired people find it challenging to understand lip movements without training, making it challenging for them to hear spoken words. They are able to interpret speech and continue social interactions without relying on aural perception thanks to audio visual speech recognition.

In order to help deaf people and provide them a method to understand what is being spoken to them, lip reading and audio-visual speech recognition was developed. It aids deaf persons in understanding spoken language words to enable them to participate actively in conversations.

The main objectives of the proposed works are (a) Create a custom database for Kannada Language, (b) Find out the lip localization method, and (c) Develop an algorithm for visual speech recognition.

This section discussed various exciting work that has been done, such as pre-processing, feature extraction, recognition methods, and accuracy, and dataset creation are discussed. Lip reading as an unaccompanied a base for communication is largely popular for the hearing impaired, and when enhanced by audio signals, it permits effortless sympathetic of speech in extremely trained subjects.

Radha N et al. proposed visual speech recognition for lip shape and lip actions to form a combined system with the combination of feature and model level. The proposed technique indicates the progress in the performance of 85% in Motion History Image—discrete wavelet transform built features, 74% in Motion History Image—discrete cosine transform, and 80% in Motion history Image-Zernike moments-based features [1].

Adriana Fernandez-Lopez et al. proposed a character-based audio-visual speech recognition for the TCD-TIMIT dataset and increased the accuracy by 10% [2]. Movellan et al. proposed a hidden Markov model for very minimum vocabulary and recognition for the first four English digits, which are one, two, three, and four. They worked on the autonomous speech recognition system, and here, they used the first four digits of English as data, this work was used for the possible application of car phone dialling, and a hidden Markov model used for recognition [3].

Stavros Petridis et al. proposed visual speech recognition for the minimum number of datasets using LSTM, and the authors suggested future extraction and classification stages, and two streams: the first one features extraction from mouth regions and the second one is extracting the features from altered images [4].

Alexandros Koumparoulis et al. introduce two contributions: the first one is a very perfect and capable lip reading model called MobiLipNetV3, and the second one is an innovative recognition model called the MultiRate Ensemble (MRE). MRE is seventy-three times more capable matched to residual neural network-based lip reading and twice as capable as MobiLipNetV2 [5]. Shridhara et al. are working on the Kannada language of the Karnataka region. In order to create a prosodically oriented phonetic search engine, the authors gather voice data in the Kannada language and describe the issues raised in the transcript. Three models—read, conversion, and extempore—are revealed in the speech corpus. A four-layered transcription, which includes IPA symbols, syllabification, pitch pattern, and pattern break, is a phonetic transcription of all the data. It is suggested that the HTK paradigm be used to create a Kannada language recognition system [6].

Kate Saenko et al. compared the results on lip detection and classified the speech and nonspeech and evaluated the performance beside several baseline systems. Dynamic Bayesian network for recognition of various loosely coordinated streams [7]. Praveen Kumar et al. developed a Kannada speech recognition system for continuous speech with diverse noisy conditions. Here, they used around two thousand four hundred speaker's data and Kaldi toolkit used for recognition model at different phoneme stages. The authors used 80% data for training and 20% for testing using kaladi [8].

Amaresh et al. worked on lip reading for the Kannada language. Authors used a custom dataset for their research and extraction of the region of interest using edge detection algorithm called canny; for feature extraction of the lip parameters, they used Gabor convolve and grey level co-occurrence Matrix algorithm. Classification by artificial neural network [9]. Ozcan et al. used trained and not pre-trained convolutional neural network models to recognition lip-reading. AV letters Dataset used for the training and testing stages, and pre-trained models, such as AlexNet and Googlnet, were used [10]. Jing Hong et al.'s research on lip reading was helpful for hearing-impaired people. Here, they used the GRID dataset and three CNN architectures, such as Spatiotemporal CNN, CNN with recurrent units, and pre-trained CNN with recurrent units [11]. Abderrahim Mesbah et al. proposed a Hahn algorithm-based lip reading, and the authors used three types of datasets, such as AVLetters, OuluVS2, and BBC LRW datasets [12].

Jing Hong et al. develop a novel method called Generative Adversarial Network, and GAN has overcome the existing machine learning approach and unseen class samples also used for training, this method is helpful to increase the accuracy of the visual speech [13]. Yuanhang Zhang et al. inclusive study to estimate the outcome of dissimilar facial sections with state of the art technology. Here, they studied visual speech means, not only lip-reading; it contains the entire face, including the mouth, upper face, and cheeks. They worked on word level and sentence level standards with different features [14].

Hassanat explained in detail visual speech recognition, he mentioned three main steps: first one is detecting the human face, the second one is lip localization or feature extraction, and the third one is lip reading. For feature extraction geometric features, appearance, image transformed, and hybrid approaches are mentioned [15]. B Soundarya et al. proposed convolutional neural network-hidden Markov model-based lip reading. Lip reading is used for teaching hearing-impaired people to communicate with other people [16].

Juergen Luettin et al. proposed active shape models and HMM-based visual speech recognition, and obtained an accuracy of 85.42% [17]. Arun Raghavan et al. compared lip-reading on two main perceptions: first one is worked on normal person hearing in a noisy environment and another one is a hearing-impaired person, and they proved visual speech recognition is helpful in a noisy environment [18]. Shashidhar R and Sudarshan proposed the speech recognition for audio visual. In this visual speech recognition, they used LSTM network and obtained 80% accuracy [19]. Sunil et al. used an active contour model for lip tracking and lip-reading, and authors used the geometrical feature extraction model and the Hidden Markov model used as a classifier. The authors used a housing dataset and compared their results with the cuave database [20]. Joon Son Chung et al. proposed three main goals: first, they collected the dataset from TV channels and produced a dataset of millions of word occurrences, pronounced over thousands of ways to different two people [21].

Ziad Thabet et al. proposed recognition of spoken word in real-time. The author's main focus is on lip reading in an autonomous vehicle. Here, they used three classifiers, gradient boosting, SVM, and logistic regression, and the results were 64.7%, 63.5%, and 59.4%, respectively [22]. Joon Son Chung et al. contributed three steps: first one is to attain a novel huge associated training corpus that encloses profile face, the second one is the program knowledge process that can cover Syncnet, and the third one is to validate the recognition of visual speech for unseen videos. Results are compared with the Oulu VS2 dataset [23]. Amit Garg et al. used VGGNet pre-trained and applied celebrity's pictures and Google images, and for feature extraction, they used LSTM [24].

Michael wand et al. used the GRID corpus dataset, which is the publicly available dataset, and authors used LSTM and SVM as classifiers and compared the output with LSTM and SVM [25]. Karel et al. proposed a lip reading-based histogram of oriented gradients from the front face of the human. The hidden Markov model used as a classifier. The authors used two publicly available datasets, which were cuave, ouluVs, and their own dataset, TULAVD [26]. Abhishek Jha et al. used LRW and GRID datasets. CMT and WAS were used as feature extraction models [27]. Different visual speech recognition techniques

using convolutional neural networks are explained. Additionally, pre-processing methods, feature extraction techniques, and recognition methods, such as multimodal VSR, and Hybrid CNN methods, are explained [28]. Sudarshan et al. proposed the VGG16 neural network for Indian English language and obtained the accuracy of 76% [29]. Philippe et al. developed a computer aided method for feature extraction of white blood cells [30]. Dhasarathan et al. suggested deep learning method to analysis COVID-19 data using homomorphic systematic approach [31]. Zarif et al. proposed single level database approaches for preserving healthcare information [32].

2. Implementation Steps

The construction of a unique data collection, pre-processing, training testing, and validation were all included in the implementation, which was carried out in stages.

2.1. Dataset Development

An elaborate setup was used to produce a customized dataset for Kannada words. Table 1 lists the attributes of the captured videos.

Table 1. Features details.

Parameter	Value
Resolution	$1080 \times 1920\ p$
Frames/Second	60 FPS
Average Duration of Video	1–1.20 s
Average Size of Video	10 Mb

The data base was created to speed up the creation and authentication of processes used to train and evaluate lip gesture-based methods. It is a collection of videos of people speaking while narrating a static character that will be used to train and recognize lip motion outlines. It included audio and lip action data from a variety of subjects explaining the same words in different videos. The Kannada word dataset was generated in a controlled environment with noise levels kept below 20 decibels. The video samples were meticulously produced and shot at full HD resolution in a well-lit environment to obtain clear and centred frames (1080×1920). The recordings were made in a quiet, controlled setting. With 11 male and 13 female subjects ranging in age from 18 to 30 years old, this data collection contains roughly 240 video clips per person.

Applications for voice recognition and lip analysis can exploit this data collecting. The numbers regarding the quantity of the data gathering are displayed in Table 2. The Kannada word for which the data collection was created is displayed in Table 3. The dataset utilized for testing and verification was limited to 10 Kannada words because to a lack of computational resources and hardware constraints.

Table 2. Dataset.

Parameters	Language
	Kannada
Quantity of Words	05
Quantity of Subjects	10
Samples per Subject	5
Total Number of Samples	250

2.2. Parameters

Machine learning is used in the proposed research. Any basic parameters must be understood to grasp the concept of the implementation.

Table 3. Summary of VGG16 Architecture.

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	224 × 224 × 3	-	-	-
1	2 × Convolution	64	224 × 224 × 64	3 × 3	1	Relu
	Max Pooling	64	112 × 112 × 64	3 × 3	2	Relu
3	2 × Convolution	128	112 × 112 × 128	3 × 3	1	Relu
	Max Pooling	128	56 × 56 × 128	3 × 3	2	Relu
5	2 × Convolution	256	56 × 56 × 256	3 × 3	1	Relu
	Max Pooling	256	28 × 28 × 256	3 × 3	2	Relu
7	3 × Convolution	512	28 × 28 × 512	3 × 3	1	Relu
	Max Pooling	512	14 × 14 × 512	3 × 3	2	Relu
10	3 × Convolution	512	14 × 14 × 512	3 × 3	1	Relu
	Max Pooling	512	7 × 7 × 512	3 × 3	2	Relu
13	FC	-	25,088	-	-	Relu
14	FC	-	4096	-	-	Relu
15	FC	-	4096	-	-	Relu
Output	FC	-	1000	-	-	Softmax

2.2.1. Activation Function (AF)

The development of a neural network is regulated by scientific equations. Each neuron in the node is active in this function, which controls whether it is activated or not, based on the neuron's input that is connected to the model's estimate.

Additionally, AF aids in standardizing each neuron's output to fall within the range of 1 to 0 or −1 to 1. It is a mathematical "gate", as depicted in Figure 1, that separates the input that serves the current neuron from its output, which travels to the following layer.

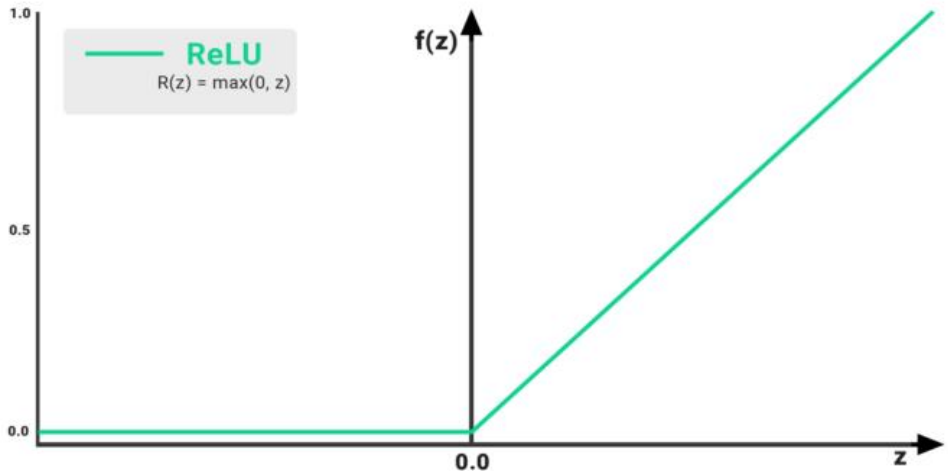

Figure 1. Role of Activation Function (AF).

As it is the most frequently, utilised AF in deep learning models, the ReLU AF was used at the input and hidden layers in this study. The function produces 0 (zero) in the case of a negative input but returns z as a linear function in the case of a positive input.

$$f(z) = R(z) = \max(0, z) \quad (1)$$

or

$$f(z) = R(z) = \begin{cases} 0 \text{ for } z \leq 0 \\ z \text{ for } z > 0 \end{cases} \quad (2)$$

The output data are the form of (0, 1), which allows a neural network to avoid binary organization and handle as many groups as possible. As a result, Softmax is also known as a multinomial logistic regression.

$$S(Z)_i = \frac{e^{zi}}{\sum_{j=1}^{k} e^{zi}} \; for \; i = 1, 2, \ldots\ldots k \quad (3)$$

2.2.2. Batch Size

The number of exercise examples utilised in a single repetition is referred to as the repetition size in machine learning. In this study, the batch size was set at 128. Figure 2 shows the overview of the activation function called Softmax.

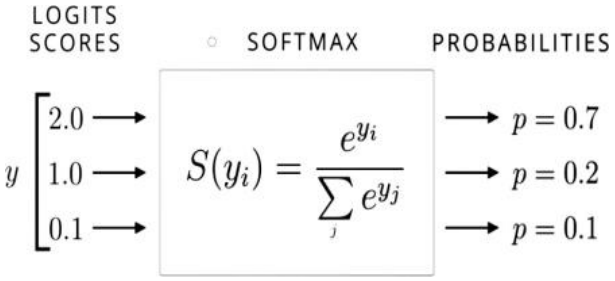

Figure 2. Softmax Activation function overview.

2.2.3. Drop Out

The majority of the restrictions in a neural network model are overcome by a fully coupled layer, and as training advances, co-dependency between later neurons develops. This causes each neuron's unique power to exceed the training data's suitability.

Dropout is a technique for achieving regularization in neural networks, which reduces inter-reliant learning and avoids over-fitting.

Training Phase: Disregard a random portion of nodes for every unseen layer, every training model, and every repetition.

Phase of testing: Utilize all activations, but scale them back by p.

The dropped nodes in a standard neural network are shown in Figure 3 to prevent overfitting and achieve regularizations.

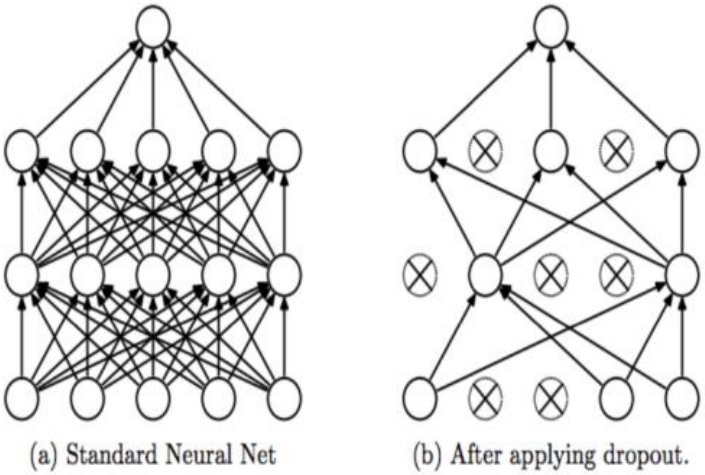

Figure 3. Representation of Dropout.

With the exception of the output layer, which had a 0.4 drop-out, all layers had drop-outs set to 0.3.

2.2.4. Cross-Entropy

It is a measurement of how different two probability distributions are from each other. The equation for Cross-Entropy can be found here:

$$I(p, q) = -\sum P(x) \log q(x) \tag{4}$$

where x represents the expected outcomes of the algorithm, $p(x)$ the probability distribution of true labels derived from training samples, and $q(x)$ the estimated outputs of the ML method.

Since there were more than two classifications, we used categorical cross-entropy. It is a loss function that is used to categorize single labels.

$$\mathrm{L}(\mathrm{y}, \hat{\mathrm{y}}) = -\sum_{j=0}^{M} \sum_{i=0}^{N} [Yij * \log(\hat{\mathrm{y}}ij)] \tag{5}$$

where y represents the real value and $\hat{\mathrm{y}}$ represents the expected value.

2.2.5. Epochs

In several passes, the machine learning algorithm has completed the entire training data set and is represented by an epoch. In most cases, data sets are grouped.

The number of epochs equals the sum of iterations if the group size is the entire training data set. In various models, more than one epoch is formed. The overall relationship between data size—d, the e—total sum of epochs, i—the total sum of iterations, and batch size—b is given by:

$$d \times e = i \times b \tag{6}$$

The data set was trained for 200–300 epochs in this study, and the difference in the Loss Function was observed for different numbers of epochs.

3. VGG16 Architecture

The VGG16 architecture is a convolutional neural network. A scalable architecture that adapts to the scale of the dataset, including its simplicity. A novel algorithm is proposed and tested with this smaller dataset in the proposed technique, with room for further progress. It is thought to be among the best vision model architectures ever created. Instead of creating a lot of hyper-parameters, stride 1 concentrates on using 3×3 filter convolution layers, and stride 2 always adjusts the same padding and total pool layer of 2×2 filters. It adheres to this convolution and max pool layer planning across the entire architecture.

It has two FC (fully connected layers) at the top, followed by a Softmax for output. The 16 in VGG16 relates to the fact that it has 16 layers, each with its own weight. This network is very large, with approximately 138 million (approximately) parameters.

The building, as depicted in Figure 4, which is based on the BBC LRW data set, uses Python code to import the weights, which are measured and saved as an image net. The Top Layer is excluded from the model in order to avoid the last few layers of the architecture. The input data is tailored to the input data stream of the $224 \times 224 \times 3$ architecture. This technique is referred to as data reshaping. The Output stage has also been condensed to a $7 \times 7 \times 512$. An example of a prediction shape is this. Due to hardware limitations, a lack of computational power, and a lack of time, researchers chose to approach the problem via transfer learning, where a previously constructed model (VGG16) is used, the weights are imported, and the model is adjusted to meet the requirements of the application. VGG16 architecture with parameters are explained in Table 3.

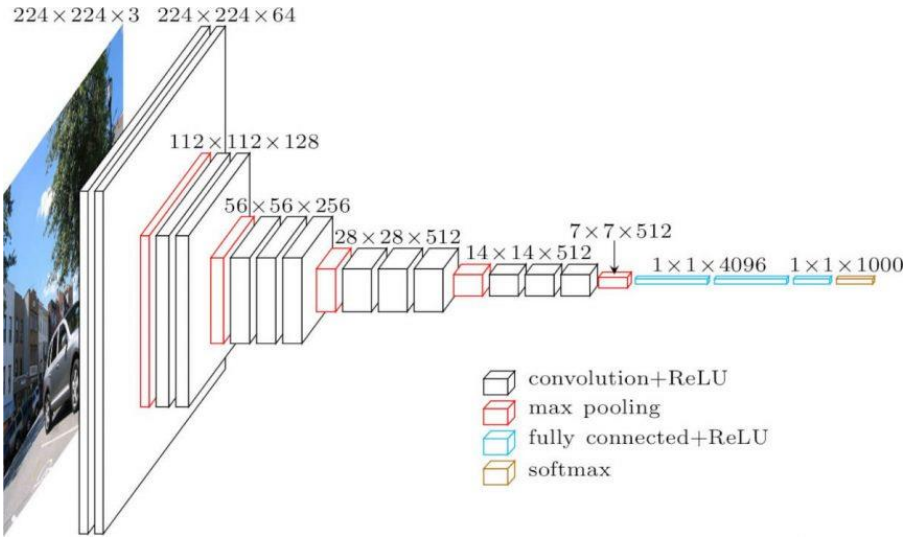

Figure 4. Architecture VGG16.

4. Result and Discussion

For Kannada dataset, a similar implementation was carried out, and the results are as follows.

The training for the video samples of five Kannada words has been completed. The teaching takes a long time to complete, and there are about 250 video clips. The data-step-by-step set's preparation tests different criteria, as seen in Figure 5. Figure 6 depicts the disparity in training and testing losses as a function to the number of epochs. Authors train the custom dataset using epochs, for this work authors used 396 epochs for training.

```
Epoch 297/300
396/396 [==============================] - 2s 5ms/step - loss: 0.2863 - accuracy: 0.8662 - val_loss: 0.2284 - val
_accuracy: 0.8889
Epoch 298/300
396/396 [==============================] - 2s 5ms/step - loss: 0.2857 - accuracy: 0.8788 - val_loss: 0.3253 - val
_accuracy: 0.8283
Epoch 299/300
396/396 [==============================] - 2s 5ms/step - loss: 0.3790 - accuracy: 0.8333 - val_loss: 0.2910 - val
_accuracy: 0.8586
Epoch 300/300
396/396 [==============================] - 2s 5ms/step - loss: 0.3636 - accuracy: 0.8712 - val_loss: 0.3071 - val
_accuracy: 0.8687
```

Figure 5. Training of Epochs for five Kannada Words.

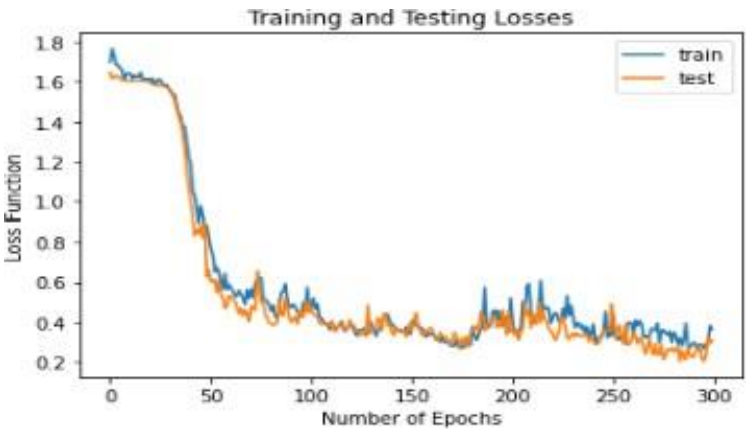

Figure 6. Number of Epochs vs. Loss function for Kannada dataset.

The altered weights are saved and loaded for Prediction after the Training is over. The difference between the prediction process and the training process is that the tag values will initially be missing, then calculated using the closest approximation method. In this method, the full data set is anticipated, and the expected tag values are gathered. They are compared to the Actual tag values to calculate the Overall Model Accuracy. The model's overall metrics or report are displayed in Table 4.

Table 4. Report on metrics for the Kannada database.

	Precision	Recall	FI-Score	Support
avanu	0.781	1000	0.877	50
bagge	1000	0.898	0.946	49
bari	0.909	1000	0.952	50
guruthu	0.980	1000	0.990	48
helidha	1000	0.700	0.824	50
accuracy			0.919	247
macro avg	0.934	0.920	0.918	247
weighted avg	0.933	0.919	0.917	247

Figure 7 plots a normalized confusion matrix to better show how the actual and predicted labels are labelled. The True/Actual Labels are represented by the *Y*-axis in the Confusion Matrix, while the Predicted are represented by the *Z*-axis. The *X*-axis serves as a symbol for labels. The Metrics Report and the Confusion Matrix show that the accuracy of the whole model for five Kannada words with 247 samples from 10 subjects is 91.90 percent. The outcomes of the video recognition paradigm are contrasted with those of other models in Table 5.

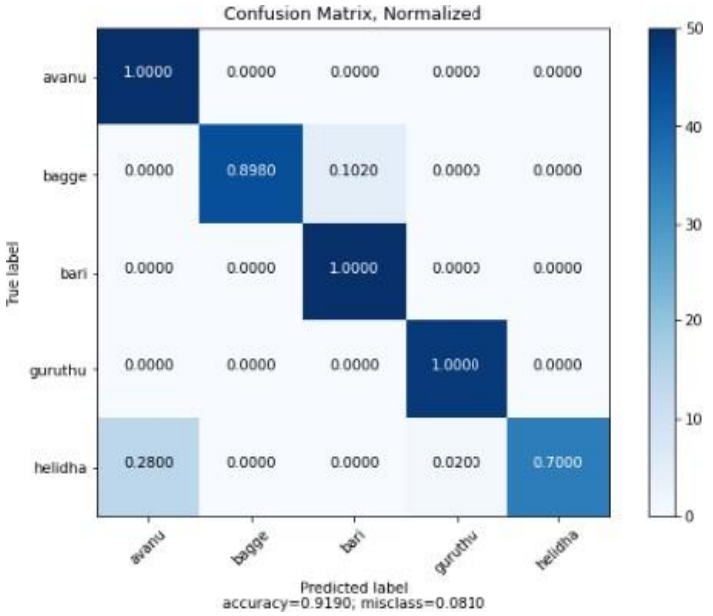

Figure 7. Confusion Matrix for Kannada Database.

The main application of the proposed method are that deaf persons can utilise visual speech recognition systems and lip-reading techniques to interpret spoken words by observing lip movements. The process of data encryption and decryption, and information security both provide appealing applications for lip reading. The limitations of the lip reading are movement of the lips are not seen many times and the geometrical region of the lip position varies from person to person.

Table 5. Comparison of Proposed Methodology with other methodologies.

Methodology	Dataset	Classes	Accuracy
HCNN (Without DA) [12]	BBC LRW	500	55.86%
HCNN (With DA) [12]	BBC LRW	500	58.02%
GLCM-ANN [9]	Custom	10 English, 10 Kannada, 10 Telugu	90.00%
VGG16 [29]	Custom	250 English	75.2%
VGG16 [Proposed]	Custom	10 Kannada = 10 × 10 × 5 = 500	91.90%

5. Conclusions

In this work, authors present visual speech recognition or lip reading for the Kannada Language. In our work, authors used our own dataset because the standard dataset is not available for the Kannada language. The proposed work used for five Kannada words, and it contains 247 samples/video clips, and video clips are converted into the number of frames and applied VGG16 CNN to train and test the datasets. The accuracy of the model is around 91.9% for Kannada words and we compared this result with existing results. In proposed method, authors used Kannada language words, other researchers may use same technique for different language datasets. In future, researchers may use different datasets or custom datasets for deep learning and machine learning algorithm to obtain better accuracy. Researchers may integrate audio and visual for better result and research may develop portable device using microcontroller and microprocessor for security and biometric authentications.

Author Contributions: Conceptualization, G.H.L., S.P.K. and S.R; data curation, S.P.K., S.R and G.H.L.; formal analysis, M.K., G.H.L. and S.R.; funding acquisition, V.R..; investigation, G.H.L., S.P.K. and S.R.; methodology, S.R., V.R. and S.P.K.; project administration, V.R., G.H.L. and M.K.; Software, G.H.L., S.P.K., M.K. and S.R, Writing—original draft, S.P.K. and S.R; Writing—review & editing, V.R.; Validation, G.H.L., S.P.K., M.K. and S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest related to this work.

References

1. Radha, N.; Shahina, A.; Khan, A.N. Visual Speech Recognition using Fusion of Motion and Geometric Features. *Procedia Comput. Sci.* **2020**, *171*, 924–933. [CrossRef]
2. Fernandez-lopez, A.; Karaali, A.; Harte, N.; Sukno, F.M. Cogans For Unsupervised Visual Speech Adaptation To New Speakers. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; Volume 2, pp. 6289–6293.
3. Movellan, J.R. Visual Speech Recognition with Stochastic Networks. *Adv. Neural Inf. Process. Syst.* **1995**, *7*, 851–858. Available online: https://papers.nips.cc/paper/1994/hash/7b13b2203029ed80337f27127a9f1d28-Abstract.html (accessed on 11 October 2022).
4. Petridis, S.; Wang, Y.; Ma, P.; Li, Z.; Pantic, M. End-to-end visual speech recognition for small-scale datasets. *Pattern Recognit. Lett.* **2020**, *131*, 421–427. [CrossRef]
5. Koumparoulis, A.; Potamianos, G.; Thomas, S.; da Silva Morais, E. Resource-adaptive deep learning for visual speech recognition. *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech* **2020**, *2020*, 3510–3514. [CrossRef]
6. Shridhara, M.V.; Banahatti, B.K.; Narthan, L.; Karjigi, V.; Kumaraswamy, R. Development of Kannada speech corpus for prosodically guided phonetic search engine. In Proceedings of the 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, India, 25–27 November 2013. [CrossRef]
7. Saenko, K.; Livescu, K.; Siracusa, M.; Wilson, K.; Glass, J.; Darrell, T. Visual speech recognition with loosely synchronized feature streams. *Proc. IEEE Int. Conf. Comput. Vis.* **2005**, *II*, 1424–1431. [CrossRef]
8. Kumar, P.S.P.; Yadava, G.T.; Jayanna, H.S. Continuous Kannada Speech Recognition System Under Degraded Condition. *Circuits Syst. Signal Process.* **2020**, *39*, 391–419. [CrossRef]

9. AKandagal, P.; Udayashankara, V. Visual Speech Recognition Based on Lip Movement for Indian Languages. *Int. J. Comput. Intell. Res.* **2017**, *13*, 2029–2041. Available online: http://www.ripublication.com (accessed on 11 October 2022).
10. Ozcan, T.; Basturk, A. Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models. *Balk. J. Electr. Comput. Eng.* **2019**, *7*, 195–201. [CrossRef]
11. Hong, J.; Nisbet, D.A.; Vlissidis, A.; Zhao, Q. *Deep Learning Methods for Lipreading*; The University of California, Berkeley Department of Electrical Engineering & Computer Sciences: Berkeley, CA, USA, 2017.
12. Mesbah, A.; Berrahou, A.; Hammouchi, H.; Berbia, H.; Qjidaa, H.; Daoudi, M. Lip reading with Hahn Convolutional Neural Networks. *Image Vis. Comput.* **2019**, *88*, 76–83. [CrossRef]
13. Kumar, Y.; Sahrawat, D.; Maheshwari, S.; Mahata, D.; Stent, A.; Yin, Y.; Shah, R.R.; Zimmermann, R. Harnessing GANs for Zero-Shot Learning of New Classes in Visual Speech Recognition. *arXiv* **2019**. [CrossRef]
14. Zhang, Y.; Yang, S.; Xiao, J.; Shan, S.; Chen, X. Can We Read Speech beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 356–363. [CrossRef]
15. Hassanat, A.B.A. Visual Speech Recognition. In *Speech and Language Technologies*; IntechOpen Limited: London, UK, 2011. [CrossRef]
16. Soundarya, B.; Krishnaraj, R.; Mythili, S. Visual Speech Recognition using Convolutional Neural Network. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1084*, 012020. [CrossRef]
17. Grewal, J.K.; Krzywinski, M.; Altman, N. Markov models—Hidden Markov models. *Nat. Methods* **2019**, *16*, 795–796. [CrossRef] [PubMed]
18. Raghavan, A.M.; Lipschitz, N.; Breen, J.T.; Samy, R.N.; Kohlberg, G.D. Visual Speech Recognition: Improving Speech Perception in Noise through Artificial Intelligence. *Otolaryngol.—Head Neck Surg.* **2020**, *163*, 771–777. [CrossRef] [PubMed]
19. Shashidhar, R.; Patilkulkarni, S.; Puneeth, S.B. Audio Visual Speech Recognition using Feed Forward Neural Network Architecture. In Proceedings of the 2020 IEEE International Conference for Innovation in Technology (INOCONF 2020), Banglore, India, 6–8 November 2020. [CrossRef]
20. Morade, S.S.; Patnaik, S. A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition. *Optik* **2014**, *125*, 5181–5186. [CrossRef]
21. Chung, J.S.; Zisserman, A. Learning to lip read words by watching videos. *Comput. Vis. Image Underst.* **2018**, *173*, 76–85. [CrossRef]
22. Thabet, Z.; Nabih, A.; Azmi, K.; Samy, Y.; Khoriba, G.; Elshehaly, M. Lipreading using a comparative machine learning approach. In Proceedings of the 2018 First International Workshop on Deep and Representation Learning (IWDRL), Cairo, Egypt, 29 March 2018; pp. 19–25. [CrossRef]
23. Chung, J.S.; Zisserman, A. Lip reading in profile. In Proceedings of the British Machine Vision Conference 2017, London, UK, 4–7 September 2017; pp. 1–11. [CrossRef]
24. Garg, A.; Noyola, J. Lip Reading Using CNN and LSTM. 2016; Available online: http://cs231n.stanford.edu/reports/2016/pdfs/217_Report.pdf (accessed on 11 October 2022).
25. Wand, M.; Koutník, J.; Schmidhuber, J. Lipreading With Long Short-Term Memory. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6115–6119. [CrossRef]
26. Paleček, K. Lipreading using spatiotemporal histogram of oriented gradients. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 28 August–2 September 2016; pp. 1882–1885. [CrossRef]
27. Jha, A.; Namboodiri, V.P.; Jawahar, C.V. Word Spotting in Silent Lip Videos. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 150–159. [CrossRef]
28. Sooraj, V.; Hardhik, M.; Murthy, N.S.; Sandesh, C.; Shashidhar, R. Lip-Reading Techniques: A Review. *Int. J. Sci. Technol. Res.* **2020**, *9*, 4378–4383.
29. Patilkulkarni, S. Visual speech recognition for small scale dataset using VGG16 convolution neural network. *Multimed Tools Appl.* **2021**, *80*, 28941–28952. [CrossRef]
30. Saade, P.; Jammal, R.E.; Hayek, S.E.; Zeid, J.A.; Falou, O.; Azar, D. Computer-aided Detection of White Blood Cells Using Geometric Features and Color. In Proceedings of the 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 20–22 December 2018; pp. 142–145. [CrossRef]
31. Dhasarathan, C.; Hasan, M.K.; Islam, S.; Abdullah, S.; Mokhtar, U.A.; Javed, A.R.; Goundar, S. COVID-19 health data analysis and personal data preserving: A homomorphic privacy enforcement approach. *Comput Commun.* **2023**, *199*, 87–97. [CrossRef] [PubMed]
32. El Zarif, O.; Haraty, R.A. Toward information preservation in healthcare systems. In *Innovation in Health Informatics, A Smart Healthcare Primer*; Academic Press: Cambridge, MA, USA, 2020; pp. 163–185. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

View publication stats