Visual Speech Recognition using VGG16 Convolutional Neural Network

Shashidhar R (shashidhar.r@sjce.ac.in)
JSS Science and Technology University,Sri Jayachamarajendra College of Engineering
https://orcid.org/0000-0002-3737-7819

S Patilkulkarni
JSS Science and Technology University

Nishanth S Murthy
JSS Science and Technology University

Research Article

Keywords: Visual Speech Recognition (VSR), Machine learning, VGG16, Convolutional Neural Networks (CNN),

Posted Date: March 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-177220/v1

Wireless Personal Communication (2021)
https://doi.org/10.1007/s00034-017-0623-5

Visual Speech Recognition using VGG16 Convolutional Neural Network

Shashidhar R[1], S Patilkulkarni[2], [3]Nishanth S Murthy

Abstract Communication is all about expressing one's thoughts to another person through speech and facial expressions. But for people with hearing impairment, it is difficult to communicate without any assistance. In most of these cases Visual speech recognition (VSR) systems simplify the tasks by using Machine Learning algorithms and assisting them to understand speech and socialize without depending on the auditory perception. Thus, one can visualize VSR system as a lifeline for people with hearing impairment which helps them in providing a way to understand the words that are being tried to convey to them through speech. In this work we used VGG16 convolutional neural network architecture for Kannada and English datasets. We used custom dataset for the research work and got the accuracy of 90.10% for English database and 91.90% for Kannada database.

Keywords Visual Speech Recognition (VSR); Machine learning; VGG16; Convolutional Neural Networks (CNN);

Shashidhar R
shashidhar.r@sjce.ac.in

Patilkulkarni
sudarshan_pk@sjce.ac.in

Nishanth S Murthy
nishanthsmurthy24@gmail.com

1 Department of Electronics and Communication Engineering, JSS Science and Technology University, Sri Jayachamarajendra college of Engineering, Mysuru, India-570006

2 Department of Electronics and Communication Engineering, JSS Science and Technology University, Sri Jayachamarajendra college of Engineering, Mysuru, India-570006

3 Department of Electronics and Communication Engineering, JSS Science and Technology University, Sri Jayachamarajendra college of Engineering, Mysuru, India-570006

1. Introduction

One recent trend in the image processing domain is Pattern Recognition. It has become an important approach by virtue of which human brain imitation and interpretation can be achieved using computers. The existing approach such as fingerprint, gesture or facial recognition has various shortcomings. These can be overcome by employing visual speech recognition, makes it more beneficial and robust which makes it an important building block of Human-machine interface. To implement pattern recognition successfully, computer vision and image processing are important steps in visual speech recognition.

During the last few decades, automated speech recognition methods were designed but the noise effects reduce the performance of these methods drastically. Due to outstanding innovations in semiconductor technology, demand for internet utility is increasing. To cater for the consumer needs, cost effective visual sensors and faster signal processing is essential.

The visual data consists of speech videos like music, news, video calls etc. For text documents, efficient systems are existing but for videos it is not so. Since for video information meta-data is required, makes the system more expensive. Speech video contains spoken speech which will be corrupted easily with noise in the media (channel). This leads to poor quality of video for speech processing. From audio-visual data, video speech can be modeled by lip motions. This is achieved in three phases like lip reading, lip synchronization and lip landmark localization. Lip reading can be described as a skill used to determine a person's words of speech by spotting the lip movements lacking the perception of sound. Hearing impaired people find it difficult to interpret the lip movements, unless they are specifically trained to do so and thereby it is challenging for them to detect the spoken words.

Lip reading involves modeling lip video clips into phonemes or characters using deep learning models. The video clips will have speaker face and lip movement. This will provide better performance than the classical methods. Since lip reading is based on recognition model it has constraints like quality of video, speaker head variations and fixed vocabulary size suffers from various performance issues. The great source of infotainment consists of movie dialogues, public speech and so on. A simple audio dubbing makes video unnatural. Hence, cross language speech dependent lip synchronization is preferred.

Lip-landmark localization involves structural representation of lips and is crucial in improving the performances of the system. The challenges are facial hair and occlusion due to microphone or hand during conversation.

2. Literature Survey

An extensive literature survey has been conducted prior to the beginning of the proposed work and have been well documented for further reference.

Vital step in speech recognition is Lip extraction from the video source or the dataset, ensuring a high recognition rate. In Active Appearance Model (AAM), the shape and grey-level appearance can be determined [1]. The aim was to extract lip areas directly, because numerous additional portions such as eyes, eyebrow, moustache and body were reflected in the target image. This method was used to extract the lip regions. This model gives the idea of how lip extraction is done in order to interpret the characters. The face region is first extracted, and then the region of interest (ROI) is customized to extract the lip region. Further, Hidden Markov Model (HMM) as well Dynamic Programming (DP) matching methods were applied, both the methods showed high recognition accuracy.

Further, up gradation to the existing system was carried out by performing the analysis of lip extraction in real time [2]. The lip movements were captured using the camera and database was created. This method operates in two modes: registration mode and recognition mode. Here, in the automation processes, primarily automatic spoken section extraction and camera control to decrease the amount of operations. To distinguish the shapes, the threshold time is set. In the camera control method, the camera is used in order to extract the captured image. The region extraction is not applied in the initial mode. The rectangular area of 80×80 pixels at the middle of a 320×240-pixel image is taken and the extracted rectangle region is used.

The Lip-reading analysis was implemented for English letters as pronounced from Filipino orators using image analysis [3]. MATLAB was used to process and format the video data gathered into a sequence of images using integrated JAV. The videotape was converted to sequences of images for the analysis. 12 image frames were taken for processing. Finally, the images were converted in *.jpg* format. Lip detection and extraction is then performed using Viola-Jones procedure and point plotting by means of Point distribution model tracking KLT Algorithm.

Active contour models or Snakes were used for shape analysis and object detection using deformable templates [4]. The extracted target contour is trans- formed into energy minimization to make it optically fit. The pixel color, intensity, corners and edges are the features extracted by the image-based detection method. This is called as color-based methods because of color difference between face as well lip. In RGB Model individual components transformed and filtered using HPF and converted into binary image to recognize the lip [4]. The hue value difference between lip pixel and face pixel is used as a criterion to recognize the lip in HSV model. In $Y\,C_bCr$ model, the differences in blue and red chroma component are used as a fact to locate the lip. Lips have more red pixels compared to faces and have high C_r and low C_b values.

A novel lip-reading algorithm [5] is proposed, which uses localized Active Contour Model (ACM), geometric parameter extraction followed by classification by HMM model. Variations in height, width and area of lip used as a feature vectors and dynamic information are captured.

Distinct features were compared and it was found that changes in vertical path of lip have substantial impact on recognition rate. The outcomes obtained via HMM with CUAVE database are relatively better than custom developed databases. The percentage recognition rate of female candidates is more than that of male candidates with an increase from 1% to 2%.

Lip movement analysis via deep neural network using hybrid visual features [6] was proposed using DBN-HMM hybrid models. Highly discriminative visual features were extracted using efficiently developed processing blocks. The application of designed Deep Belief Network (DBN) based recognizer is emphasized. Multi-speaker (MS) and orator-independent (OI) tasks performed over CUAVE database and phoneme recognition rates (PRRs) 77.65% and 73.40% were obtained respectively. The finest word recognition rates realized in the tasks of MS and OI 80.25% and 76.91% respectively. This method overcomes all disadvantages faced by the conventional Hidden Markov Model.

An appearance-based feature extraction process was pro- posed which introduced Deep Belief Network (DBN) based recognizer [6]. It showed better performance than HMM baseline recognizer. Visual based features were extracted in the automatic speech recognition system to give a baseline accuracy of 29.8%. Using visual features as inputs resulted in best DBN architecture achieving an accuracy of 45.63

AAM is a hybrid method, combining both model-based and pixel-based methods. The advantage of this model is being able to distinguish the words from whatsoever angle of the extracted lip pictures. It describes the gray-level variation of an object with a set of model parameters to detect the lip The set of labeled landmark points are taken as a parameter to define the shape of the object. x and y coordinates are used to locate each land- marked point. Principal component analysis (PCA) is used for building statistical shape models by taking a training library of a land marked object in images. The shape of an object deviated from the mean shape is detected from the Eigen vectors and Eigen values of a covariance matrix.

A lip-reading system using HMM where DCT and DWT was proposed. It was based on features extracted from the mouth region and compared with DCT and DWT based features [8]. HMM with DWT based features gave good results with 97% performance when compared to HMM with DCT which gave only 91%. The main objective of this paper was to improve communication between a normal person and hearing- impaired person.

Lip movement analysis methodology based on 3-D DCT and 3-D HMM is proposed, based on a 3-dimensional approach [9]. It offers good robustness to the performance of conventional 1-D DCT and 1-D HMM, in a way that it accommodates changes to rotation, parallel shift and scaling of the test sub- ject. This method slightly increases the recognition rate about 2-3% against the conventional method.

A hybrid lip reading technique using Convolutional Neural Networks and Long Short-Term [10] was proposed. The words and phrases were predicted from VGG net pre-trained video samples of human faces of celebrities from IMDB and Google images, without the

use of audio signals. They achieved a validation accuracy of about 76%. As

well achieved a 47.57% success rate for an upstretched model and up to 59.73% success rate for stretched model. The LSTM model took a long time to train, especially while updating the VGGNet and it does not handle the sequence until feature extraction is complete.

An advanced technique using CNN and Bi-directional Long Short-term Memory [11] was proposed using the Caffe toolbox and Tensor flow toolbox. It was claimed that this method outperformed conventional methods like Active Con- tour model (ACM) and HMM.A Lip-reading technique based on HMM and Cascade feature extraction. [12] Was proposed. Viola-Jones method was used and the algorithm was applied for detection of Chinese characters composed of training and testing phrases. The four- Cascade feature extraction and HMM was proposed. Viola-Jones' approach was used and an algorithm was applied for detection of Chinese characters composed of training and testing phrases. The four-stage cascaded method included DCT and DWT based image transformation, PCA based dimensional reduction, K-means based vector quantification and HMM based recognition. The DCT-PCA method yielded an outcome of 72.8% when the characteristic vector has a dimension of 35 and the involvement rate of the particular eigen values is 98%. The DWT-PCA method yielded a result of 77.4% with a dimension of 42 when the involvement rate of the selected Eigen values is 97%.

Visual speech recognition as a speaker-dependent problem was described [13]. The inference was drawn by comparing the word error rates (WER) of both speaker-dependent as well as speaker-independent experiments. It was found to be 76.38% and 33% respectively. Speaker dependent experiments gave better results than speaker independent experiments. Charlie Chaplin videos were used and the main aim was to spot the words in silent talking without implicitly identifying the spoken words, in which lip motion of the orator was clearly observable and audio was absent. The authors developed a pipeline for identification-free salvage, and show its performance in contradiction of recognition-based reclamation on a significant number of dataset and one more set of out-of-vocabulary words. The word spotting process achieves 35% increased mean average precision over identified-based methods on a wide range LRW dataset. Validate the application of the technique by word noticing in a prevalent speech video [14].

A canny edge detection algorithm was proposed for extraction of region of interest and for feature extraction Gray Level Co-occurrence Matrix and Gabor convolve algorithm was used. The classification was implemented using artificial neural networks which attained an accuracy of 90% [15]. Different views of the speaker were used for lip reading using a pose normalization block in a standard system. The effects of pose normalization on the audio-visual integration strategy are analyzed by AV-ASR [16].

Publicly available data called GRID corpus was used and Lip reading was achieved successfully by replacing visual speech recognition pipeline with com- pact neural network architecture. Feature extraction was done using HMM; later LSTM architecture was used and accuracy of 79.6% [17] was achieved. Lip-reading is to find what the speakers say by the movement of lip only. Proposed model is composed of 3D Convolutional layered with Dense Net and residual bidirectional long short-term memory.

Hanyu Pinyin (a phonemic transcription of Chinese) was used as a tag and entirely had 349 classes, although the number of Chinese letterings is 1705 [18]. Based on histogram of oriented gradients, visual speech parameterization was proposed for lip reading and integration based on HMM and as a classification algorithm which got 89.9% accuracy after fusion of some parameterizations via multi-stream synchronous [19]. A machine learning approach was developed to recognize lip reading using a benchmark dataset which consists of one million words. Nine classifiers were used, among those three got the best result namely Support vector machine (SVM), Logistic regression (LR) and Gradient Boosting as 63.5%, 59.4% and 64.7% respectively [20].

A 95.2% accuracy was achieved using GRID corpus database and they pro- posed according to authors LipNET it is the foremost end-to-end sentence-level lip-reading prototype [21]. A five hundred AR face database was used for implementation in MATLAB. The author proposed a limited active contour model-based technique used to segment the lip area. Lip separation is essential to graphic lip-reading systems, because the precision of segmentation result directly affects the recognition rate [22]. a model called Watch, Listen, Attend and spell commonly called as WLAS model was contributed. The WLAS model trained the LRS dataset which consists of 100000 natural sentences from British television. The LRW dataset was trained using WAS model and a 23.8%-word error rate was achieved. For GRID dataset, the word error rate was 3.0% [23].

Lips can read in profile features but this standard is inferior to frontal faces. A new large aligned corps MV-LRS was obtained, that contain profile faces selected using a face pose repressor network with the accuracy of 88.9% [24]. Words were recognized only by video in the absence of audio using continuous speech. CNNs were used to investigate individual words for direct recognition. CNN and LSTM architecture got excellent results which are used to classify temporal lip motion sequence of words [25].

Feature improvements techniques to reduce speaker variability were examined where HMM was used for recognition. In this work low level, image-based features were compared with high-level model-based features for lip reading. The two approaches were investigated for correcting the speaker dependence of the visual features: namely per-speaker z-score normalization and Hi-LDA [26]. Three methods were proposed for VSR, the first one is using the HMM model to recognize the image sequences, the second one is top-down approaches which used a principal component analysis for lip-reading features. The third one is a bottom up approach that uses a nonlinear scale space scrutiny to form structures straight from the pixel intensity. The AV letter database was used for implementation [27].

Around ninety data were collected from five subjects and six isolated words three times. A grid-based feature extraction was used for some isolated words an accuracy of 60% was achieved [28]. Photographic features are mostly classified into shape based and appearance based. A new set of hybrid visual features was proposed which lead to an improved pictorial speech recognition system. Pseudo-Zernike Moment is considered for shape-based visual feature while Local Binary Pattern-three orthogonal planes and Discrete Cosine Trans- form are considered for the appearance-

based feature. Artificial Neural Network (ANN), multiclass Support Vector Machine (SVM) and Naive Bayes (NB) distinguishers were implied for classifier hybridization [28].

A technique for extraction of features was proposed called spatiotemporal discrete cosine transform. For individual and combined classification Support vector machines and tailor – made Hidden Markov models were used respectively [29].

The current trends in visual speech recognition were evaluated and it was shown that pictorial speech plays an important role in Automatic Speech Recognition and also Authors discussed different type's databases [30].The aim of proposed work is to predict the spoken word, given video of a person speaking in the absence of audio and vice versa. It has been carried out at various phases like Develop a database for English Kannada language, Develop an algorithm for Video recognition, validate the data under the trained system and achieve the best performance.

3. Proposed Methodology

The objectives are then converted in a detailed design flow to make the task of VSR simple. The Fig 1 represents the block diagram of the proposed visual speech recognition system.

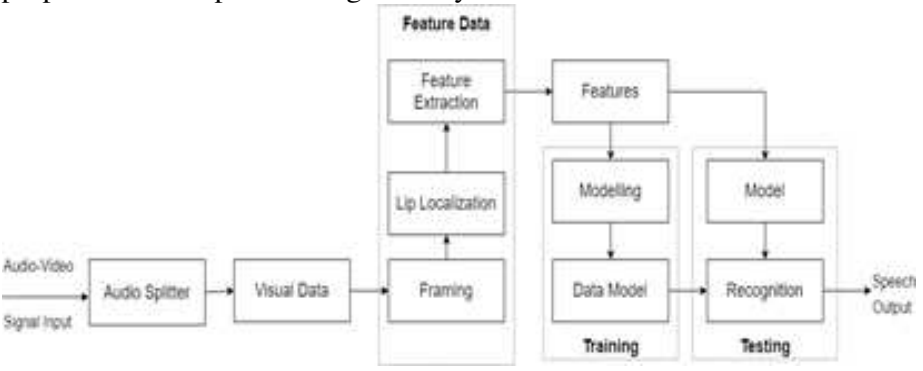

Figure 1: Block Diagram of Proposed Method

It can be seen that input audio-video signal is split into audio and video channels, only the Video data is taken for processing. The visual video recognition has several important steps namely pre-processing, feature extraction, training-testing and recognition using convolutional neural networks system.

3.1. Hardware Requirements

The proposed methodology includes training of large data-sets followed by testing and validation of test sample, all of which require robust and efficient processing units.

This work was conducted on a PC equipped with an Intel Core i7 processor, 8th Generation CPU, assisted along with RAM of 8GB to handle complex ML and AI algorithms. A storage space of 100 GB was utilized on whole to store the huge volume of data-set and also for the

execution of the developed models.

The requirement of video camera for recording video and a Microphone to record the audio was fulfilled by a Smartphone and an Electronic Gimbal to produce stable videos.

3.2. Software Requirements

This work was executed using the Ubuntu 18.04.5 LTS - Bionic Beaver (as a guest OS) upon Oracle VirtualBox VM. The Jupyter Notebook Environment was used to execute the pre-processing and machine learning methods, algorithms with the Python 3.8 command set. Being an open-source web application it allows to create and share documents containing live code, equations, visualizations and narrative text. It is widely used for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning etc.

Adopting the block-by-block execution of code sets allows rapid development, easy debugging and visualization options. All the necessary libraries can be imported in the run window. Some of the important libraries include,

- **OpenCV** : designed to solve computer vision problems
- **Numpy**: general-purpose array-processing package
- **Scipy**: high level data-manipulation and data-visualization
- **Matplotlib**: Python 2D plotting library
- **dlib**: toolkit containing machine learning algorithms and tools
- **Keras**: open-source neural-network library
- **Scikit-learn**: free software machine learning library
- **pandas**: data manipulation and analysis

4. Implementation

The implementation was carried out in a step-by-step process that involved creation of custom data-set, pre-processing, training testing and validation.

4.1. Data-set Creation

Custom data-set was created for both English and Kannada Words using an extensive setup. The features of the recorded videos are as shown in Table 1.

Table 1: Data-set Features

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

The creation of the data-set was done to facilitate the development and validation of algorithms used to train and test the system that consists of lip- motion. It is a collection of videos of volunteers reciting a fixed script that is intended to be used to train software to recognize lip-motion

patterns. It comprised of correlated audio and lip movement data in multiple videos ofmultiple subjects reading the same words.

The dataset of spoken English and Kannada words was created in a controlled environment and it was made sure that noise levels were maintained well below 20db. The video samples were carefully composed and shot in a well-lit area at Full HD resolution (1080 X 1920) in order to obtain a sharp and focused frames. The recordings were collected in a controlled, noise-free, indoor setting. This data-set consists of around 240 video samples per person, 11 male and 13 female subjects, with ages ranging from 18yrs to 30yrs, volunteered for the data-set creation process.

This data-set can be used for speech recognition and lip reading applications. The numerical values for the quantity of the data-set are as shown in Table 2.

The English and Kannada word for which the data-set was created is illustrated in Table 3.

The dataset used for the testing and verification was limited to 10 English and 10. Kannada words due to the lack of computational resources and hardware limitations.

Table 2: Data-set

Parameters	Language	
	English	Kannada
Number of Words	20	27
Number of Subjects	24	24
Samples per Subject	5	5
Total Number of Samples	2400	3240

Table 3: List of English and Kannada words

Sl No.	English Words	Kannada Words
1	Part	ಹೋಗು
2	Name	ನಾಳೆ
3	Good	ಕನ್ನಡ
4	Read	ಗುರುತು
5	Where	ಹೌದು
6	Bad	ಇಲ್ಲ
7	Come	ಕೇಳಿದ
8	Dog	ನೀನು
9	Cow	ನೀವು
10	Book	ನಾನು
11	Number	ಸಭೆ
12	People	ಅವನು
13	Water	ನನಗೆ
14	Today	ಜನರಿಗೆ
15	About	ಮರಳಿ
16	English	ಬರಿ
17	Bottle	ಓದು
18	Mobile	ನೂಕು
19	Pencil	ಮಾತು
20	Paper	ಮಠ
21		ವಿಭಾಗ
22		ಕಥೆ
23		ಮಾನ್ಯತೆ
24		ಮನ್ನಣೆ
25		ಬಗ್ಗೆ
26		ತಳ್ಳು
27		ಮೈಸೂರು

4.2 *Parameters*

The proposed work makes use of machine learning. For better understanding the concept of the implementation, it is necessary to understand some of basics parameters.

4.2.1 *Activation Function*

Activation functions are mathematical equations that determine the output of a neural network. The function is attached to each neuron in the network, and determines whether it should be activated ("fired") or not, based on whether each neuron's input is relevant for the model's prediction. Activation functions also help normalize the output of each neuron to a range between 1 and 0 or between -1 and 1. It is a mathematical "gate" in between the input feeding the current neuron and its output going to the next layer as shown in Fig 2

Figure 2: Role of Activation Function

In this work the ReLU activation function was used at the input and hidden layers as it is the most commonly used activation function in deep learning models. The function simply outputs the value of 0 if it receives any negative input, but for any positive value z, it returns that value back like a linear function.

$$f(z) = R(z) = max(0, z)$$

(or)

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

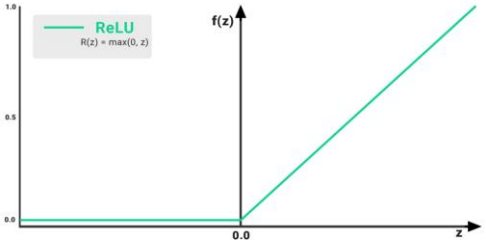

Figure 3: Performance of ReLU Activation Function

The ReLU function as shown in Fig 3 is non-linear and is able to back-propagate the errors and have multiple layers of neurons. ReLU takes care of several problems faced by the Sigmoid and the Tanh, hence was quickly adopted. Another activation function used at the output layer is the softmax. The Softmax regression is a form of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1. The output values are between the range [0,1] through which it is possible to avoid binary classification and accommodate as many classes or dimensions in a neural network model. Hence, softmax is often referred to as a multinomial logistic regression.

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

4.2.2 Batch Size

Batch size is a term used in machine learning and refers to the number of training examples utilized in one iteration. In this work the batch size was set to 128.

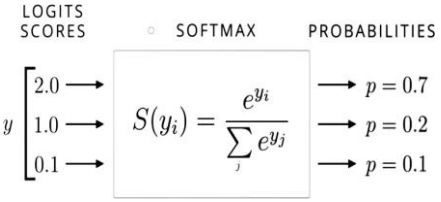

Figure 4: Overview of softmax Activation Function

4.2.3 Drop-out

Fully connected layer occupies most of the parameters in a neural network model and hence neurons develop co-dependency amongst each other during training. This curbs the individual power of each neuron leading to over fitting of training data.

Dropout is an approach to achieve regularization in neural networks thus reducing interdependent learning amongst the neurons and prevents over-fitting. It does this by adding a penalty to the loss function. By adding this penalty, the model is trained such that it does not learn interdependent set of features weights.

- **Training Phase**: For each hidden layer, for each training sample, for each iteration, ignore (zero out) a random fraction p, of nodes (and corresponding activations).
- **Testing Phase**: Use all activations, but reduce them by a factor p (to account for the missing activations during training).

Fig 5 illustrates the dropped nodes in a standard neural network to avoid over fitting and achieving regularization.

Some of the observations about dropouts are

1. Dropout forces a neural network to learn more robust features that are useful in conjunction with different random subsets of the other neurons.

2. Dropout roughly doubles the number of iterations required to converge. However, training time for each epoch is less.

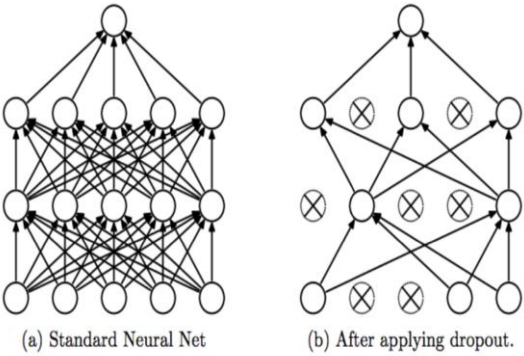

Figure 5: Representation of Drop out

3. With H hidden units, each of which can be dropped, there are 2^H possible models. In testing phase, the entire network is considered and each activation is reduced by a factor p.

Drop-outs were set to 0.3 to all the layers except the output layer for which the drop-out was set to 0.4.

4.2.4 Cross-Entropy

Entropy is defined as the smallest average size of the encoding per transportation by which any source can send data efficiently to the destination with- out any loss of information.

Cross-entropy is a measure of the degree of dissimilarities between two probability distribution. In connection with supervised machine learning, one of the probability distributions shows the label "true" for training samples and correct replies are indicated with the value hundred percent.

Cross-Entropy can express by the equation

$$I(p.q) = -\Sigma p(x) \log q(x) \tag{4}$$

Where x represents the predicted results by ML algorithm, $p(x)$ is the probability distribution of "true" label from training samples and $q(x)$ depicts the estimation of the ML algorithm.

In this work categorical cross-entropy was used as the numbers of classifications were more than 2. It is a loss function that is used for single label categorization. This is when only one category is applicable for each data point. In other words, an example can belong to one class only.

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

Where y is actual value and $\hat{\mathrm{y}}$ is the predicted value.

Categorical cross-entropy will compare the distribution of the predictions (the activations in the output layer, one for each class) with the true distribution, where the probability of the true class is set to 1 and 0 for the other classes. To put it in a different way, the true class is represented as a one-hot encoded vector and the closer the model's outputs are to that vector, the lower the loss.

4.2.5 Epochs

An epoch indicates the number of passes of the entire training data-set the machine learning algorithm has completed. Data-sets are usually grouped into batches (especially when the amount of data is very large).
If the batch size is whole training data-set then the number of epochs is the number of iterations. For practical reasons, this is usually not the case. Many models are created with more than one epoch. The general relation where data- set size (d), number of epochs (e), number of iterations (i), and batch size (b) is given by

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

In this work the data-set was trained for 200-300 epochs and the variation of Loss Function was observed for varying number of epochs.

4.2.6 Optimizer

The role of an optimizer is to update the weight parameters, to minimize the loss function. Loss function acts as guides to the terrain telling the optimizer if it is moving in the right direction to reach the bottom of the valley, i.e., the global minimum.
In this work Adam Optimizer was used. It calculates the individual adaptive learning rate for each parameter from estimates of first and second moments of the gradients is the Adam function.
It is computationally more efficient and has very little memory requirement compared to its counterparts. Hence it is one of the most popular gradient descent optimization algorithms.
Adam algorithm first updates the exponential moving averages of the gra- dient (m_t) and the squared gradient (υ_t) which is the estimates of the first and second moment. Hyper-parameters $\beta 1, \beta 2 \in [0, 1)$ control the exponential decay rates of these moving averages as shown below

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{7}$$

$$\upsilon_t = \beta_2 \upsilon_{t-1} + (1 - \beta_2) g_t^2 \tag{8}$$

Where, m_t and v_t are the estimates of first and second moment respectively.

Moving averages are initialized as 0 leading to moment estimates that are biased around 0 especially during the initial time-steps. This initialization bias can be easily counteracted resulting in bias-corrected estimates.

$$\hat{m}t = \frac{mt}{1-\beta_1^t} \tag{9}$$

$$\widehat{vt} = \frac{Vt}{1-\beta_2^t} \tag{10}$$

Where $\widehat{m}t$ and vt are the bias corrected estimates of first and second moment respectively.

Finally, the parameter is updated as shown below

$$\theta_{t+1} = \theta_t - \frac{n\widehat{m}_t}{\sqrt{\widehat{v}_t}+\varepsilon} \tag{11}$$

4.3 VGG16 Architecture

VGG16 is a convolution neural net (CNN) architecture. Though very simple is a versatile architecture that adapts itself to the size of the dataset. In the proposed methodology, an novel algorithm is put forward and is verified with this smaller dataset with the scope for further developments (increase in the dataset used).
It is considered to be one of the excellent vision model architecture till date. Instead of having a large number of hyper-parameter, it focuses on having convolution layers of 3×3 filter with a stride 1 and always uses same padding and max pool layer of 2×2 filter of stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end it has 2 FC (fully connected layers) followed by a softmax for output. The 16 in VGG16 refers to it has 16 layers that have weights. This network is a pretty large network and it has about 138 million (approx) parameters.

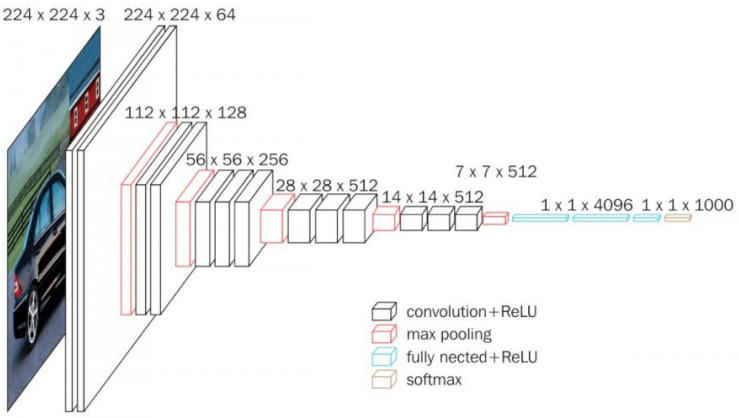

Figure 6: VGG16 Architecture

The Architecture as shown in Fig. 6 is built around the BBC LRW data-set, the weights are calculated and stored as *imagenet* which is imported through the python code. The Top Layer is removed in the model to eliminate the last few layers in the architecture.
The input data is configured to suit the input data-stream for the

224×224×3 architecture. This step is called as Data Re-Shaping. Similarly, the Output stage is also streamlined to have a shape of $7 \times 7 \times 512$. This is called as the prediction shape. Bound with hardware limitations, computational resources and time we de- cided to approach the problem using transfer learning, where in a previously created model (VGG16) can be utilized, the weights can be imported and then the model can be fine-tuned to meet the application requirements.

5 Results and Analysis

The results of the implementation of Machine Learning CNN algorithm is discussed in two sections: one for English Words and other for Kannada Words

5.1 English Words

The training for the 10 English Words video samples is executed. The training takes long time as it has around 500 video samples. The step-by-step train- ing of the data-set evaluates various parameters like Training Loss, Training Accuracy, Validation loss and Validation Accuracy as shown in Fig 7.

```
Epoch 297/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3548 -
accuracy: 0.8733 - val_loss: 0.3402 - val_accuracy: 0.8450
Epoch 298/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3475 -
accuracy: 0.8758 - val_loss: 0.3857 - val_accuracy: 0.8300
Epoch 299/300
797/797 [==============================] - 4s 5ms/step - loss: 0.3157 -
accuracy: 0.8770 - val_loss: 0.3503 - val_accuracy: 0.8550
Epoch 300/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3014 -
accuracy: 0.8808 - val_loss: 0.3374 - val_accuracy: 0.8600
```

Figure 7: Training of Epochs for 10 English Words

The Variation of the Training and Testing Loss with respect to the number of Epochs is as shown in Fig 8.

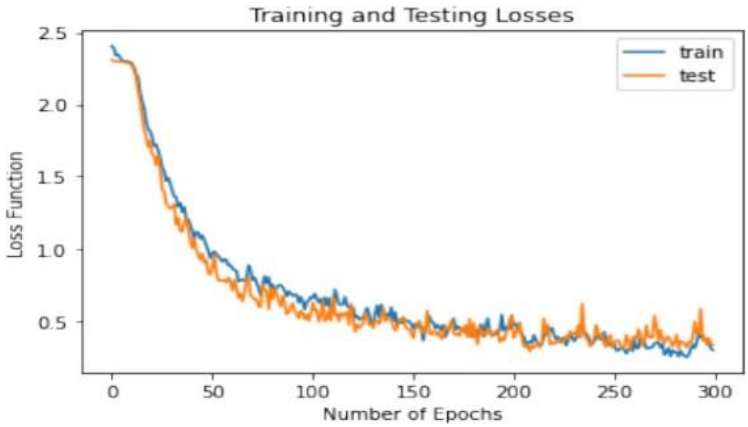

Figure 8: Variation of Training and Testing loss with Number of Epochs for English Data-Set

After the Training is completed, the updated weights are saved and loaded for the purpose of Prediction.

The Process of Prediction involves the same process as that of training, but the only difference is that the tag values will be absent initially and will be later predicted based on the closest approximation method.

In this implementation the prediction is carried out for the entire data-set and the Predicted tag values are obtained, which are then compared with the Actual tag values and the Overall Model Accuracy is calculated. The Metrics or Report of the Entire model is as shown in Table 4.

For the sake of better understanding about the Actual and Predicted Labels, a Normalized Confusion Matrix is plotted and is as shown in Fig 9. The Confusion Matrix is a Diagonal Matrix where in the Y-axis refers to the True/Actual Labels and the X-axis refers to the Predicted Labels. It can be observed from the Metrics Report and the Confusion Matrix, the Accuracy of the Entire Model is 90.10% for 10 English Words with 495 samples from 10 subjects.

Table 4: Metrics Report for English data-set

	precision	**recall**	**fi-score**	**support**
about	0.882	0.900	0.891	50
book	0.694	1000	0.820	50
come	0.980	1000	0.990	50
english	0.889	0.960	0.923	50
mobile	0.889	0.960	0.923	50
name	0.891	1000	0.942	49
pencil	0.978	0.880	0.926	50
read	1000	0.860	0.925	50
today	0.974	0.740	0.841	50
water	1000	0.696	0.821	46
accuracy			**0.901**	495
macro avg	0.918	0.900	0.900	495
weighted avg	0.917	0.901	0.901	495

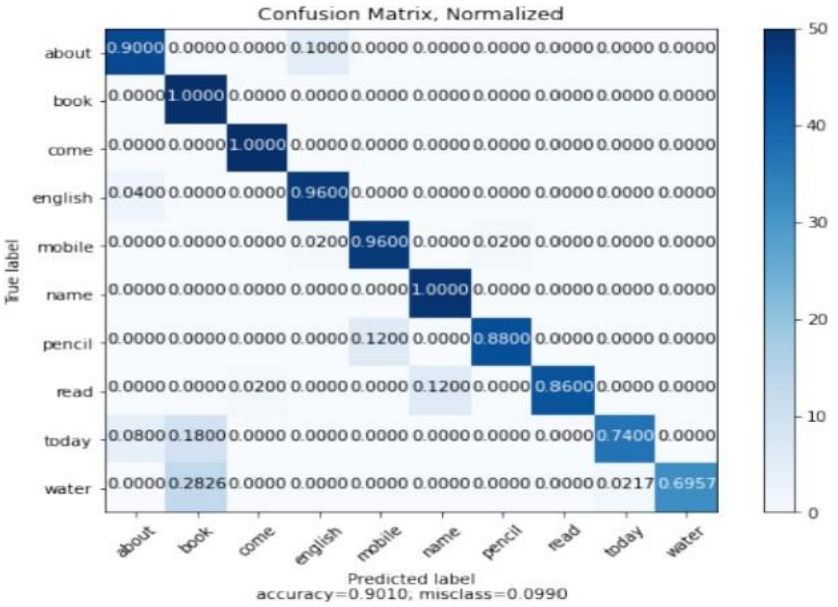

Figure 9: Normalized Confusion Matrix for English Data-Set

5.2 Kannada Words

The similar implementation was carried out for the Kannada data-set and the results are as shown.

The Training for the 5 Kannada Words video samples is executed. The training takes a long time as it has around 250 video samples. The step-by-step training of the data-set evaluates various parameters like Training Loss, Training Accuracy, Validation loss and Validation Accuracy as shown in Fig 10.

```
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 10: Training of Epochs for 5 Kannada Words

The Variation of the Training and Testing Loss with respect to the number of Epochs is as shown in Fig 11.

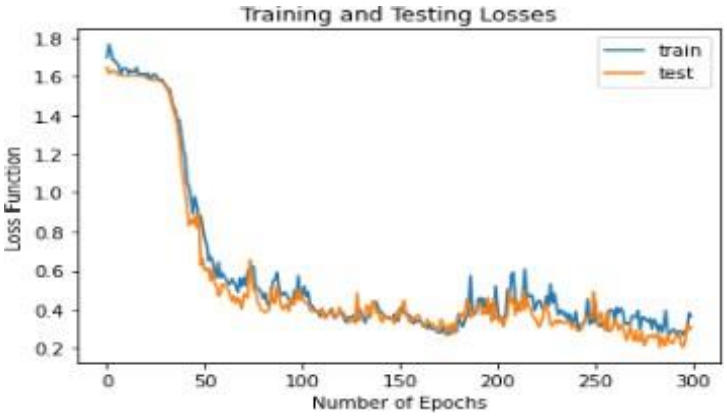

Figure 11: Variation of Training and Testing loss with Number of Epochs for Kannada Data-Set

After the Training is completed, the updated weights are saved and loaded for the purpose of Prediction. The Process of Prediction involves same process as that of training, but the only difference is that the tag values will be absent initially and will be later predicted based on the closest approximation method. In this implementation the prediction is carried out for the entire data-set and the Predicted tag values are obtained. They are compared with the Actual tag values and the Overall Model Accuracy is calculated. The Metrics or Report of the Entire model is as shown in Table 5.

For the sake of better understanding about the Actual and Predicted Labels, a Normalized Confusion Matrix is plotted and is as shown in Fig 12. The Confusion Matrix is a Diagonal Matrix where in the Y-axis refers to the

True/Actual Labels and the X-axis refers to the Predicted Labels. It can be observed from the Metrics Report and the Confusion Matrix, the Accuracy of the Entire Model is **91.90%** for 5 Kannada Words with 247 samples from 10 Subjects. The results obtained from the video recognition model are compared with the results from other methods and the comparison is shown in Table 6.

Table 5: Metrics Report for Kannada data-set

	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

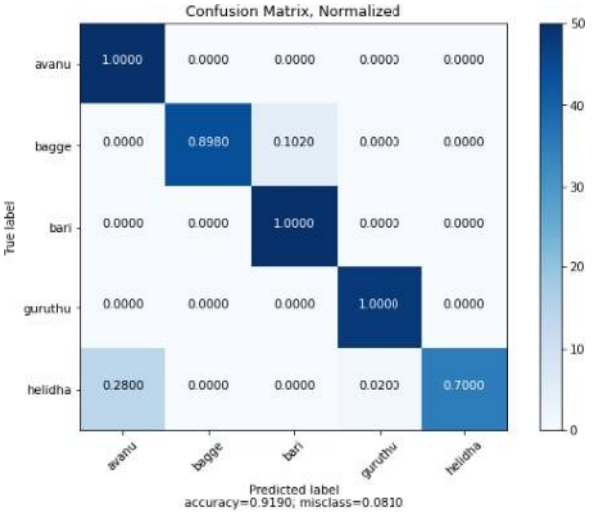

Figure 12: Normalized Confusion Matrix for Kannada Data-Set

Table 6: Comparison of Proposed Methodology with other methodologies

Methodology	Data-set Used	Classes	Accuracy		
			Top 1%	Top 5%	Top 10%
HCNN (Without DA)	BBC LRW	500	55.86%	82.93%	89.95%
HCNN (With DA)	BBC LRW	500	58.02%	84.54%	90.86%
ResNet-LSTM	BBC LRW	500	82.97%	96.28%	98.3%
Bi-LSTM & ResNet	BBC LRW	500	88.08%	96.28%	-
GLCM-ANN	Custom	10 English 10 Kannada 10 Telugu	90%		
Proposed CNN (for Video)	Custom	10 English	90.10%		
		5 Kannada	91.90%		

6 Conclusion and Future Scope

A simplified machine learning algorithm has been proposed for visual recognition of speech signals using convolutional neural networks. All the objectives that were formulated were approached systematically and completed to the fullest. Custom data-set for Kannada and English language words is created and a visual speech recognition model to interpret visual data is built for English and Kannada language words are performed separately.From the results, it is clear that even with a simplified approach high performance can be obtained which are evaluated and compared with the performance of previous methodologies and implementations. In conclusion, it can be seen that the proposed methodology does out-perform other existing methodologies. The proposed model can be easily scaled for larger data-set with more number of words irrespective of language. The proposed methodology is easily compatible with other algorithms and hence is versatile in nature. With the addition of feature extraction algorithms such as Facial Landmarks can further enhance the performance of the model.

Lip-synchronization involving synchronization of lip movement with voice (audio) can increase the cost of the system. With the integration of Audio recognition a hybrid model can be developed to increase the validation accuracy. Real time challenges are currently on the scope for future implementation. Further optimization can be attempted to decrease the training-testing and validation times, and can be deployed on a portable device such as RaspberryPi, widens the scope for authentication, authorization and security.

Declarations

Funding:

The authors did not receive support from any organization for the submitted work.

Conflicts of interest/Competing interests:

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from shashidhar.r@sjce.ac.in

Availability of data and material: Not Applicable

Code availability: Not Applicable

References

[1] T. Saitoh, R. Konishi, A study of influence of word lip reading by change of frame rate, in: Audio Visual Speech Processing (AVSP), 2010.

[2] T. Saitoh, Development of communication support system using lip read- ing, IEEJ Transactions on Electrical and Electronic Engineering 8 (6) (2013) 574–579. doi:10.1002/tee.21898.

[3] H. M. Cruz, J. K. T. Puente, C. Santos, A. V. Larry, R. Vairavan, Lip reading analysis of english letters as pronounced by filipino speakers using image analysis, Journal of Physics: Conference Series 1019 (2018) 012041. doi:10.1088/1742-6596/1019/1/012041.

[4] A. Brahme, U. Bhadade, Lip detection and lip geometric feature extrac- tion using constrained local model for spoken language identification using visual speech recognition, Indian Journal of Science and Technology 9 (32). doi:10.17485/ijst/2016/v9i32/98737.

[5] S. S. Morade, S. Patnaik, A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition, Optik 125 (18) (2014) 5181–5186. doi:10.1016/j.ijleo.2014.05.011.

[6] F. Vakhshiteh, F. Almasganj, A. Nickabadi, Lip-reading via deep neural networks using hybrid visual features, Image Analysis & Stereology 37 (2) (2018) 159–171. doi:10.5566/ias.1859.

[7] Y. Lu, J. Yan, K. Gu, Review on automatic lip reading techniques, Inter- national Journal of Pattern Recognition and Artificial Intelligence 32 (07)

(2018) 1856007. doi:10.1142/S0218001418560074.

[8] N. Puviarasan, S. Palanivel, Lip reading of hearing impaired persons using hmm, Expert Systems with Applications 38 (4) (2011) 4477–4481.

[9] K. Y. Min, L. H. Zuo, A lip reading method based on 3-d DCT and 3-d HMM, in: Proceedings of 2011 International Conference on Electronics and Optoelectronics, IEEE, 2011. doi:10.1109/ICEOE.2011.6013060.

[10] A. Garg, J. Noyola, S. Bagadia, Lip reading using CNN and LSTM, 2016.

[11] Y. Lu, J. Yan, Automatic lip reading using convolution neural network and bidirectional long short-term memory, International Journal of Pattern Recognition and Artificial Intelligencedoi:10.1142/S0218001420540038.

[12] D. Wu, Q. Ruan, Lip reading based on cascade feature extraction and HMM, in: 2014 12th International Conference on Signal Processing (ICSP), IEEE, 2014. doi:10.1109/ICOSP.2014.7015211.

[13] A. B. A. Hassanat, Speech and Language Technologies, InTechOpen, 2011, Ch. 14, pp. 279–302.

[14] A. Jha, V. P. Namboodiri, C. V. Jawahar, Word spotting in silent lip videos, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018. doi:10.1109/wacv.2018.00023.

[15] A. P. Kandagal, V. Udayashankara, Visual speech recognition based on lip movement for indian languages, International Journal of Computational Intelligence Research 13 (8) (2017) 2029–2041.

[16] V. Estellers, J.-P. Thiran, Multi-pose lipreading and audio-visual speech recognition, EURASIP Journal on Advances in Signal Processing 2012 (1). doi:10.1186/1687-6180-2012-51.

[17] M. Wand, J. Koutnik, J. Schmidhuber, Lipreading with long short-term memory, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016. doi:10.1109/icassp.2016. 7472852.

[18] X. Chen, J. Du, H. Zhang, Lipreading with DenseNet and resBi-LSTM, Signal, Image and Video Processing 14 (5) (2020) 981–989. doi:10.1007/ s11760-019-01630-1.

[19] K. Palecek, Lipreading using spatiotemporal histogram of oriented gradi- ents, in: 2016 24th European Signal Processing Conference (EUSIPCO), IEEE, 2016. doi:10.1109/eusipco.2016.7760575.

[20] Z. Thabet, A. Nabih, K. Azmi, Y. Samy, G. Khoriba, M. Elshehaly, Lipreading using a comparative machine learning approach, in: 2018 First International Workshop on Deep and Representation Learning (IWDRL), IEEE, 2018. doi:10.1109/iwdrl.2018.8358210.

[21] Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, Lipnet: End-to- end sentence-level lipreadingarXiv:1611.01599.

[22] Y. Lu, Q. Liu, Lip segmentation using automatic selected initial contours based on localized active contour model, EURASIP Journal on Image and Video Processing 2018 (1). doi:10.1186/s13640-017-0243-9.

[23] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, Lip reading sentences in the wild, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017. doi:10.1109/CVPR.2017.367.

[24] J. S. Son, A. Zisserman, Lip reading in profile, in: Procedings of the British Machine Vision Conference 2017, British Machine Vision Association, 2017. doi:10.5244/c.31.155.

[25] J. S. Chung, A. Zisserman, Learning to lip read words by watching videos, Computer Vision and Image Understanding 173 (2018) 76–85. doi:10. 1016/j.cviu.2018.02.001.

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

[27] Y.-K. Kim, J. G. Lim, M.-H. Kim, Comparison of lip image feature extrac- tion methods for improvement of isolated word recognition rate, Science & Engineering Research Support soCiety, 2015. doi:10.14257/astl.2015. 107.14.

[28] S. Debnath, P. Roy, Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition, Signal, Image and Video Processingdoi:10.1007/s11760-020-01717-0.

[29] T. Heidenreich, M. W. Spratling, A three-dimensional approach to visual speech recognition using discrete cosine transformsarXiv:1609.01932.

[30] Z. Zhou, G. Zhao, X. Hong, M. Pietikäinen, A review of recent advances in visual speech decoding, Image and Vision Computing 32 (9) (2014) 590– 605. doi:10.1016/j.imavis.2014.06.004

Figures

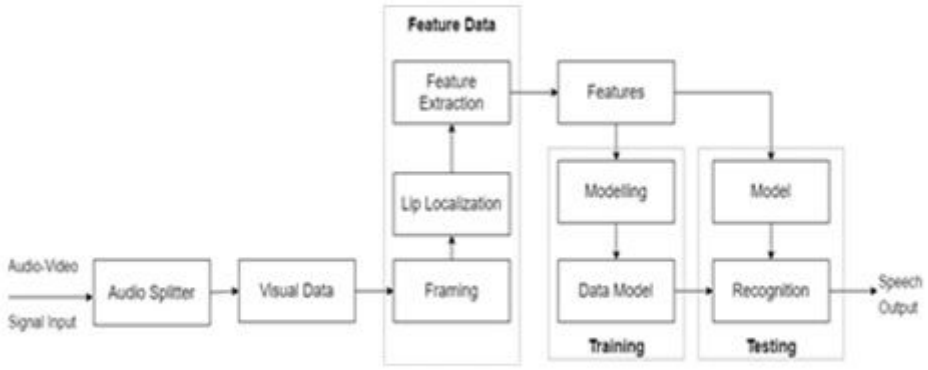

Figure 1

Block Diagram of Proposed Method

Figure 2

Role of Activation Function

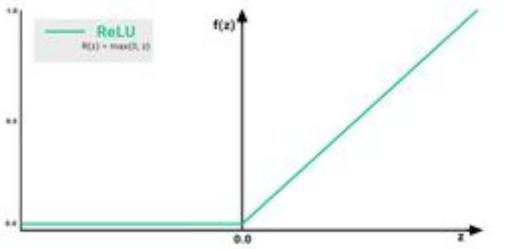

Figure 3

Performance of ReLU Activation Function

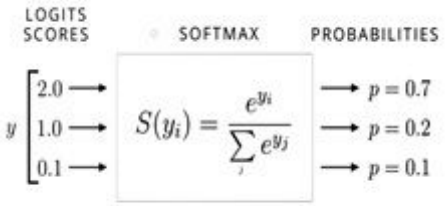

Figure 4

Overview of softmax Activation Function

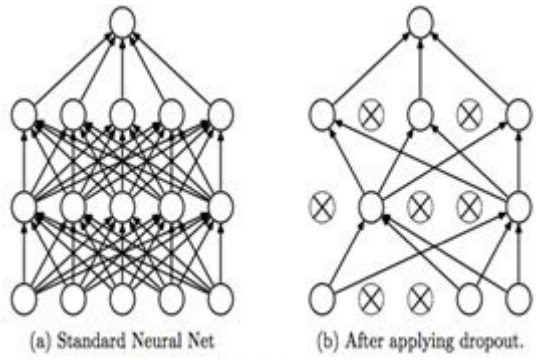

Figure 5

Representation of Drop out

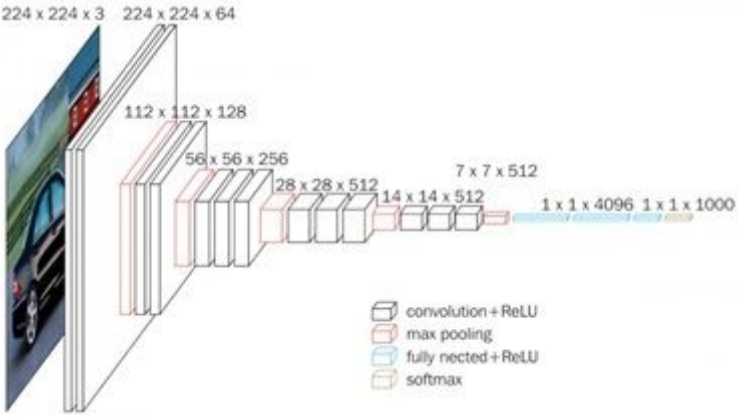

Figure 6

VGG16 Architecture

```
Epoch 297/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3548 -
accuracy: 0.8733 - val_loss: 0.3402 - val_accuracy: 0.8450
Epoch 298/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3475 -
accuracy: 0.8758 - val_loss: 0.3857 - val_accuracy: 0.8300
Epoch 299/300
797/797 [==============================] - 4s 5ms/step - loss: 0.3157 -
accuracy: 0.8770 - val_loss: 0.3503 - val_accuracy: 0.8550
Epoch 300/300
797/797 [==============================] - 4s 6ms/step - loss: 0.3014 -
accuracy: 0.8808 - val_loss: 0.3374 - val_accuracy: 0.8600
```

Figure 7

Training of Epochs for 10 English Words

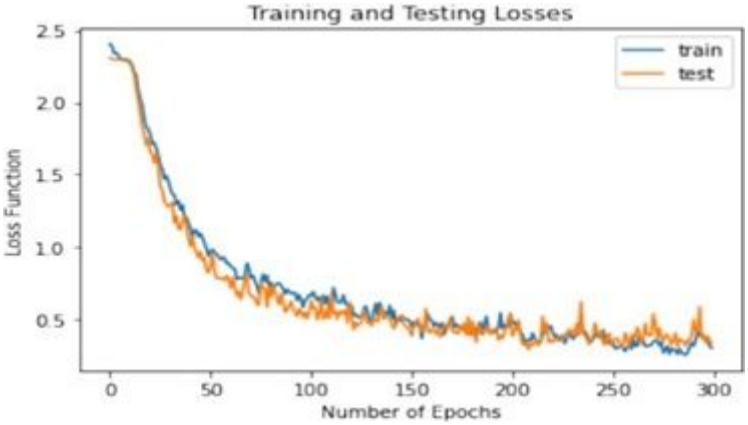

Figure 8

Variation of Training and Testing loss with Number of Epochs for English Data-Set

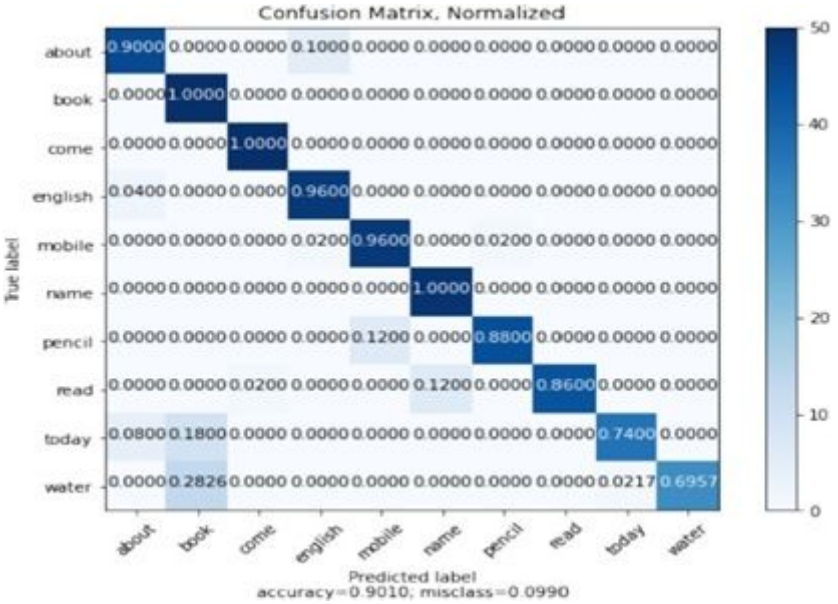

Figure 9

Normalized Confusion Matrix for English Data-Set

```
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 10

Training of Epochs for 5 Kannada Words

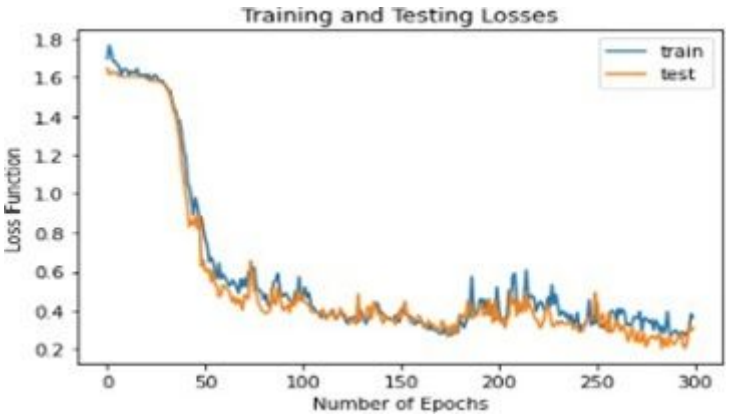

Figure 11

Variation of Training and Testing loss with Number of Epochs for Kannada Data- Set

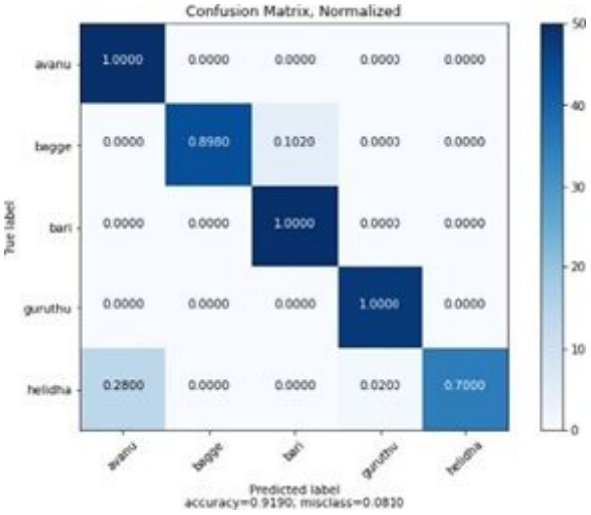

Figure 12

Normalized Confusion Matrix for Kannada Data-Set