SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V10I1P109

Volume 10 Issue 1, 90-105, January 2023

Original Article

SEOA DRN: Social Exponential Optimization Algorithm Based Deep Residual Network for Visual Speech Recognition

G. N. Srikanth[1], M. K. Venkatesha[2]

[1]*Department of EIE, RNS Institute of Technology, Bengaluru, Karnataka, India.*
[2]*RNS Institute of Technology, Bengaluru, Karnataka, India.*

[1]*Corresponding Author : srikanthgn27@gmail.com*

Received: 29 November 2022 Revised: 06 January 2023 Accepted: 16 January 2023 Published: 29 January 2023

Abstract *- The recognition of visual speech is considered an emerging solution for feasible recognition. However, the choice of imperative features is a challenging task for acquiring elevated performance. A deep model is devised for lip reading-based visual speech recognition. The CFPNet-M is being employed for extracting the regions of lips. The Lipreading technique helps to provide a silent interface and enhances speech recognition in noisy platforms, as the optical signal is not impacted via noise. The features, like Convolutional Neural Network (CNN) features, Gabor features, width, area, mass, location, orientation, Local Gabor Ternary Pattern (LGTP), statistical features, along with the voice features and spectral features, are considered. With the aid of a deep residual network (DRN), speech recognition is carried out effectively, wherein weight update of DRN is performed using Social Exponential Optimization Algorithm (SEOA). The resultant output of SEOA-based DRN is considered for visual speech recognition. The experimentation of the proposed method is done using certain measures by illustrating the efficiency of each technique. The proposed SEOA-DRN offered high efficiency with elevated accuracy of 88.4%, sensitivity of 90.6% and specificity of 90.6%.*

Keywords *- Visual speech recognition, Deep Residual Network, Voice signals, Video frames, Lip reading.*

1. Introduction

The recognition of lipreading is a procedure of detecting speech by managing lip movement in which the audio is mistreated. The preliminary techniques in this domain are extracting the features using the mouth's interesting region and trying to model its dynamics for detecting speech. The Lipreading models can facilitate the utilization of quiet borders and improve recognition of acoustic speech as visual signal is not influenced by noise [8,10]. The process of lip reading is adapted at word, alphabet and sentence levels [11]. Static images and time series classification techniques are utilized in alphabet-based lip reading.

Meanwhile, the classical models are favoured in word and sentence-based functions. The sound and image-based features can be utilized for reading lips. Specifically, the data comprises image-based features that pose an elevated success rate in applications in which they are utilized. The success rate of lip reading relies on the classification model utilized along with selecting features [7]. The major issue in detecting the speech based on video recordings is individual utterances segmentation, which helps to detect words from the frames considering video. In audiovisual speech recognition methods, speech segmentation is attained throughout the audio signals with transliterated video quantity. In some cases, the audio signals are not accessible or extremely infected by noise [21,28].

The reading with lip aimed to detect speech by understanding lip movement [12], and it is a method utilized by several people suffering from hearing loss [13]. The vision also increases audio efficiency under unfavourable acoustical conditions [37]. Reading through lip is extensively utilized in recognizing speech, identity, human-computer interfacing and multimedia models. It poses two major units, which are the extraction of lip features and recognition of features with front and back ends. However, some lipreading models utilize deep learning techniques and end-to-end models, which manifestly describe explanatory features. Several models utilize pre-trained deep learning techniques and need highly data-intensive techniques, which can be highly time-consuming during training [15].

Another problem is a deficiency of explainability in these models. This technique utilizes image unswervingly as the contribution, with nil features that made it complex to imagine and elaborate features. Even though the current investigation examined the method which utilized CNN for self-learning [16], the deep learning techniques are very complex to process. The two-stage technique extracted features with several classical image methods, like Discrete Cosine Transform (DCT) [17] and Principal Component Analysis (PCA), which are converted to various dimensions and are not instinctive to humans when analyzed [1]. Another technique is to utilize shape or appearance technique that can be used to set mouth regions

and determine the geometric features, which takes more time for training. The features must be instinctive to comprehend the behaviour of the network [1]. Recently, deep learning methods have shown huge benefits in several domains, like the representation of the image, detection of an object, recognition of humans and recognition of speech. The CNN has made a huge achievement in representing and classifying the image. Some attempts are devised to adapt CNN for learning, representing lip movement and offering emerging outcomes are generated [8]. Various deep learning models are devised to extract features using the pixels and replace the classical feature extraction phase. Fewer techniques are devised that jointly learn mined features and carry out the classification of visual speech, which has caused novel deep-learning-based lip reading models that outperformed the classical techniques. The popular deep learning techniques need huge data for effective performance. Its success led to small databases, which are self-effacing and led scientists to declare deep methods do not execute simple tasks. Thus, classical visual speech recognition techniques are improved selection whenever huge databases are not accessible [2][30,31].

The aim is to design a method for visual speech recognition based on lip reading using a deep model. The lip region is extracted from each frame using CFPNet-M. The information on the lip is obtained with the CFPNet-M, which detects the spoken words based on patterns of movements in the lip while speaking. The tracking of lip movements of a provided speaker is done to mine pertinent speech features. In addition, the DRN is introduced to perform speech recognition. DRN is trained by developed SEOA and is developed by blending Social Optimization Algorithm (SOA) and Exponentially Weighted Moving average (EWMA). The proposed SEOA-based DRN is utilized for recognizing visual speech. Here, the features of the image and signals are considered for recognition. The optimum tuning of DRN weight is done using SEOA and is produced by the unification of SOA and EWMA. The remaining sections include: Section 2 depicts priorly devised lip reading-based visual speech recognition models. Section 3 describes the designed technique for visual speech recognition. Section 4 presents the competence of the classical technique, and section 5 presents the conclusion.

2. Related Work

Eight priorly presented visual speech recognition strategies are inspected. Xuejie Zhang *et al.* [1] developed a lightweight feature extraction technique for visual speech recognition. Here, the 3D geometric features were extracted with Gabor-based image patches. This method extracted less dimensionality lip attributes. However, the consumption of memory was very high. To reduce memory usage, Stavros Petridis *et al.* [2] developed a model based on fully-connected layers and Long-Short Memory (LSTM) networks that were apposite for small databases. This method comprises two streams wherein one extracted features using mouth images and obtained features with various images but took more training time. Hong Liu et al. [3] developed an audio visual speech recognition (AVSR) method using lip graph with bidirectional synchronous fusion for visual speech recognition to reduce training time. However, the consumption of energy was high. To reduce energy consumption, Pingchuan Ma *et al.*[4] devised a hybrid Attention model using ResNet-18 and a Convolution-augmented transformer for recognizing the visual speech. Here, the audio and visual encoders were learned for extracting the features using audio waveforms and raw pixels. The fusion was done with Multi-Layer Perceptron (MLP) for speech recognition but endured elevated computational complexity. In order to decrease computational complexity, Nilay Shrivastava *et al.* [5] developed a deep neural network (DNN) model for recognizing visual speech. Here, the accuracy and count of the parameter were effectively balanced. Moreover, the depth-wise Three-dimensional convolution was utilized with channel shuffling for recognizing the visual speech. However, the technique failed with other datasets. Wentao Yu et al. [36] designed a clear stream integration network for recognizing AV speech to consider different datasets. However, the computational cost was very high. To reduce the computational cost, T. Ozcan and A. Basturk [7] developed a self-designed CNN using lip reading to recognize visual speech. The augmented AvLetters database was utilized for the training and testing stages. Here, the tuning of the mini-batch size parameter was done to recognize the speech. However, the technique can generate false recognition. Yuanyao Lu and JieYa [8] devised a hybrid model combining Bidirectional LSTM (BiLSTM) and CNN for lip reading to perform visual speech recognition to reduce false recognition. Here, keyframes were extracted to locate the mouth region. Then, the features were mined with raw mouth images considering CNN. The time taken to process a task was more. Many issues tackled by classical lip reading, like in [4], the AV model is devised that outperformed the audio-only model at higher noise levels; therefore, the adaptive fusion technique learned to consider each modality at several noise levels. To deal with this problem, the MobiVSR method is devised [5], but since it has used fewer parameters, it has difficulty tuning hyperparameters for balancing efficiency and accuracy. To address this issue explicit stream integration network is devised [36]. However, the performance of recognition was poor when utilizing more visual data. Hence, the issue relies on involving more visual data to improve recognition performance. To enhance recognition performance, CNN is devised. However, the method used a data augmentation model, which failed to offer improved outcomes all time because of the complexity of several databases [7].

3. Proposed Modelling

The proposed model is SEOA-based DRN for lip reading-based recognition of visual speech. Lip reading is a procedure to detect what a person is saying by evaluating the visual signal from their lips. The Visual signal represents the images of the mouth that exhibits changing its position during speech.

However, it is complex as the technique needs more practice and a strong base of the speaking language. The goal is to design an optimized deep model for lip reading-based visual speech recognition. Initially, the input video with word utterance is given to the frame extraction module to convert the video into multiple frames. After that, the lip region is extracted from each frame using CFPNet-M [20]. Moreover, the relevant features, like CNN features, Gabor features [1], width, area, mass, location (x-pos, y-pos), orientation from patch identified [1], LGTP, and statistical features, are extracted from the lip region. Concurrently, the voice sample is employed as input and fed to the feature extraction phase. Here, certain features, such as MKMFCC [23], BFCC [22] and spectral features, including spectral centroid, tonal power ratio, spectral spread, pitch chroma, and spectral flux, are considered. After significant feature extraction from audio and video samples, feature concatenation is done. At last, speech recognition is done using the DRN [39]. Furthermore, the DRN is trained with SEOA. The developed SEOA approach is devised here by combining SOA [38] and EWMA [9]. The architecture of the proposed SEOA-based DRN model for visual speech recognition is revealed in figure 1.

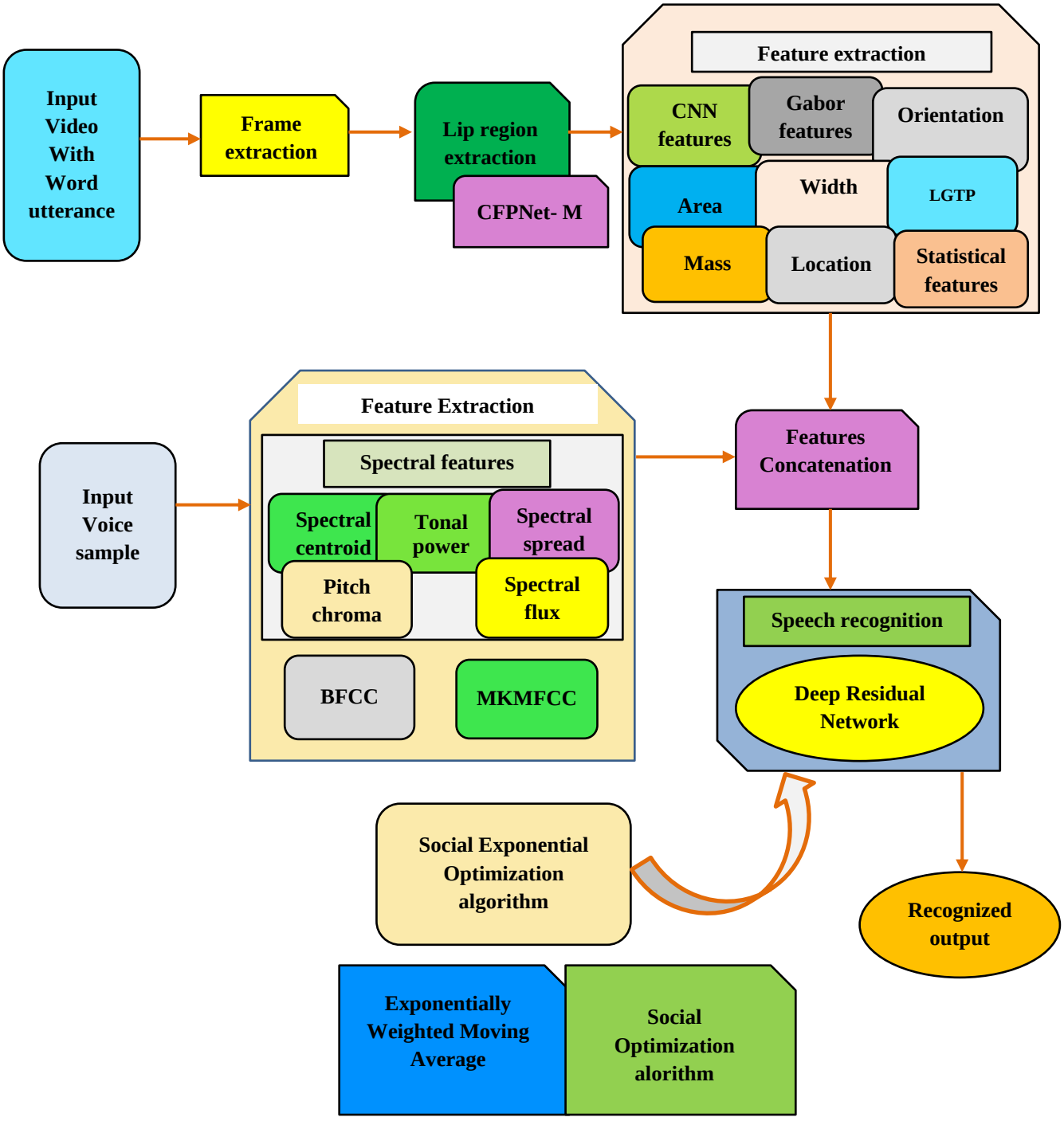


Fig. 1 Architecture of developed SEOA-based DRN for lip reading-based visual speech recognition

3.1. Acquisition of Inputs

The video samples and voice samples are acquired from the dataset for performing effective lip reading-based visual speech recognition.

3.1.1. Input Video Samples

Assume a dataset$M$that contains x videos and is expressed as

$$M = \{V_u; \quad 1 \leq u \leq x\} \tag{1}$$

where, x signifies total videos present in the database, and V_u represent u^{th}video.

3.1.2. Input Voice Samples

Consider voice signal as input which is accumulated with database$S$, and given by,

$$S = \{A_1, A_2, \ldots A_e, \ldots A_k\}; 1 \leq e \leq k \tag{2}$$

where, S refers voice signal database,$A_e$ise^{th} signal and k signify the total voice signal.

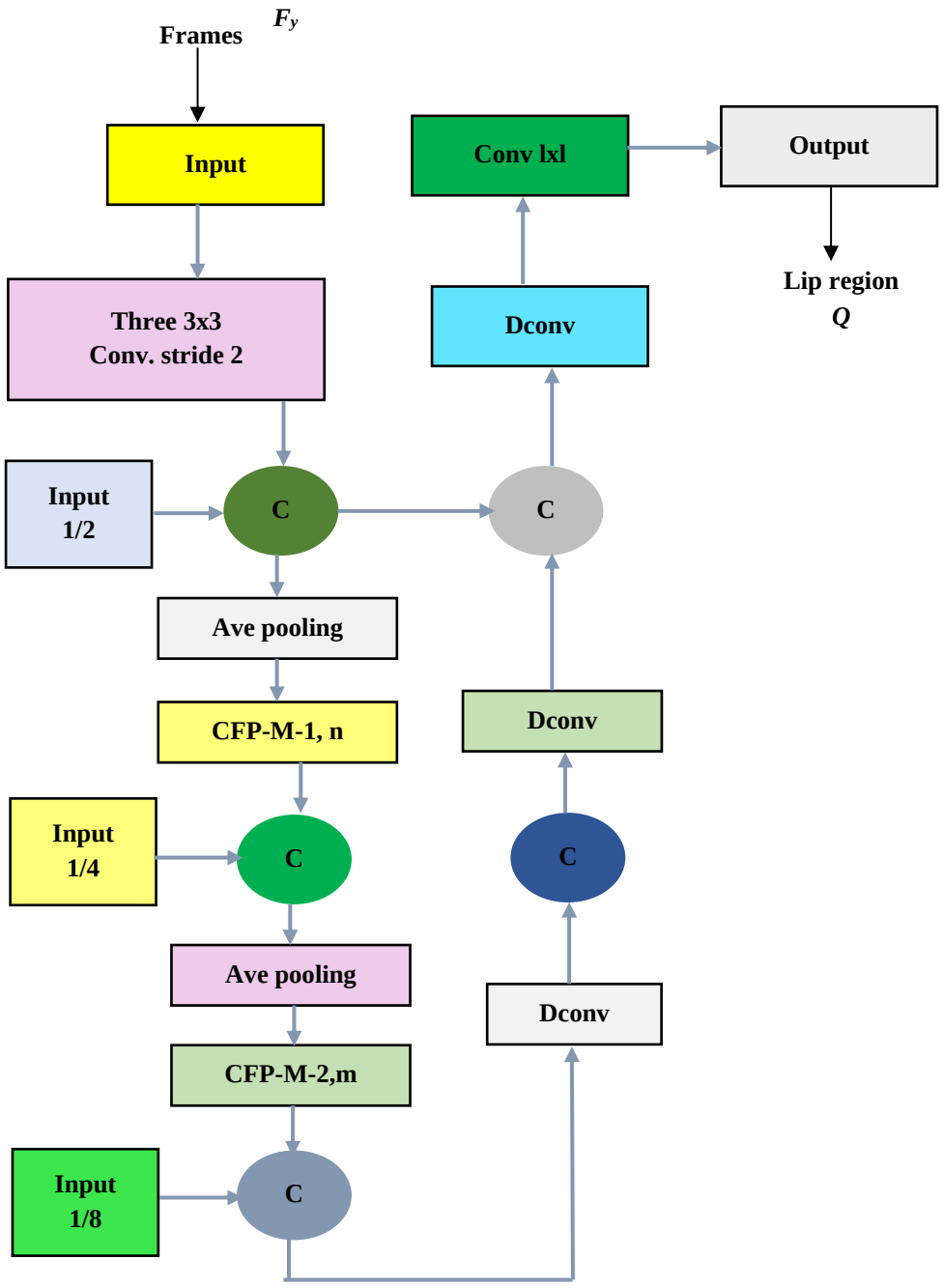


Fig. 2 CFPNet-M model

3.2. Extraction of the Video Frame

The video $V_u$is considered as frame extraction input. Extracting frames is a significant procedure that divides the video into various frames. It is useful to attain accurate data from the frames. Furthermore, it assists in preserving the salient feature and discards the frequent frames. The frames are extracted after a particular time instance. The number of frames generated from frame extraction is done using frames per second of a particular video. Here, the video with a long time contains more frames compared to videos with less duration. The count of frames is expressed as

$$\ell = f \times v \tag{3}$$

where, ℓsignifies frame count, v is video duration and $f$symbolize frames per second. Hence, each video contains specific frames, which are expressed as,

$$K = \{F_y; \quad 1 \leq y \leq o\} \tag{4}$$

where, o are total frames and F_y represent y^{th} frame.

3.3. Lip Region Extraction using CFPNet-M

The frames F_y are given as input to lip region extraction. Initially, the details regarding the CFP module are utilized for constructing the CFPNet-M. It is liable to perform real-time segmentation with enhanced accuracy. Moreover, it poses the potential to discover, quantify, and classify signal patterns. It uses fewer parameters to process a task, reducing overall computation time. Here, the count of FP channels is selected that is$K = 4$. Consider input poses size$M = 32$; for instance, the total number of filters for each channel is 8. The count of filters from the first to third convolutional operators is set to 2, 2, and 4. Subsequently, each FP channel's group of various dilation rates are set. Here, the rate of dilation is set tor_k. For instance, the first and fourth channel dilation rates are set to $r_1 = 1$andr_k. For extracting local and Gabor features, the rates of dilation of the second and third channels are set to $r_2 = r_k/4$and$r_3 = r_k/2$. Thus, CFP can study the features with medium size.

3.3.1. Structure of CFPNet-M Network

The structure of CFPNet-M is displayed in figure 2. The three 3×3 convolutional operators are utilized as the first extractor of the feature. The initial operators are utilized using stride 2 for performing down-sampling. Once the initial mining is completed, the average pooling is employed for performing down-sampling. Before other CFP (CFP-M-2), the average pooling layer is inserted for downsampling purposes. The CFP is repeated m times for building the CFP-M-2 cluster. After that, three deconvolutional operators are adapted with stride 2 for building the decoder and linking the same stage encoders by skip connections. At last, a1×1 convolution is utilized to activate the final feature map and produce the segmentation masks. In CFPNet-M, the selection of CFP module with$n = 2$, $m = 6$ and dilation rate $r_{k_{CFP-M^{-1}}} = [2,2]$ and$r_{k_{CFP-M^{-2}}} = [4,4,8,8,16,16]$. The extracted lip region is denoted as$Q$.

3.4. Attaining Imperative Features

The voice signal $A_e$is adapted as input. Concurrently, the extracted lip region $Q$is considered for extracting the image features, like CNN features, Gabor features [1], width, area, mass, location (x-pos, y-pos), orientation from patch identified [1], LGTP, and statistical features. Each feature mined using lip region is described.

3.4.1. Acquisition of Features using Frame

From the obtained lip region Q, the extraction of the image features, like CNN features, Gabor features [1], width, area, mass, location (x-pos, y-pos), orientation from patch identified [1], LGTP, and statistical features are obtained. The elaboration of all features is defined.

CNN Features

It refers to the neural network using various layers like the convolution layer, pooling (max pool) layer, and fully connected layer. The output of the convolution function is represented by,

$$D(\ell) = \left(n * W(\ell)\right) \tag{5}$$

Here, $n$symbolize input of CNN, $D(\ell)$refers to feature map, and $W(\ell)$refers kernel. The CNN feature is expressed asB_1. Figure 3 signifies CNN structure.

Gabor Features

The Gabor feature [1] is computed using the Fast Fourier Transform (FFT), which produces negative and positive along with imaginary and real units wherein the real unit is utilized. Here, each image is transformed to greyscale, and the filtering of Gabor is adapted. For minimizing the small values, like background noise and for controlling patches of image size, a threshold is adapted to the initial transform. Thus, the extracted Gabor feature is denoted asB_2.

Orientation

It is utilized for computing the orientation [1] amongst each patch. It differed from the Gabor wave orientation, which is linked to each patch orientation and is denoted as B_3.

Area

The height is modelled with Gabor wavelength, and thus area [1] is considered a good measure for mouth opening and is denoted asB_4.

Mass

Mass [1] relies on intensity, and it reveals the depth of the mouth and offers 3D representation, which can differentiate between the closed and open mouth, revealing the teeth and fully opened mouth and is denoted as B_5.

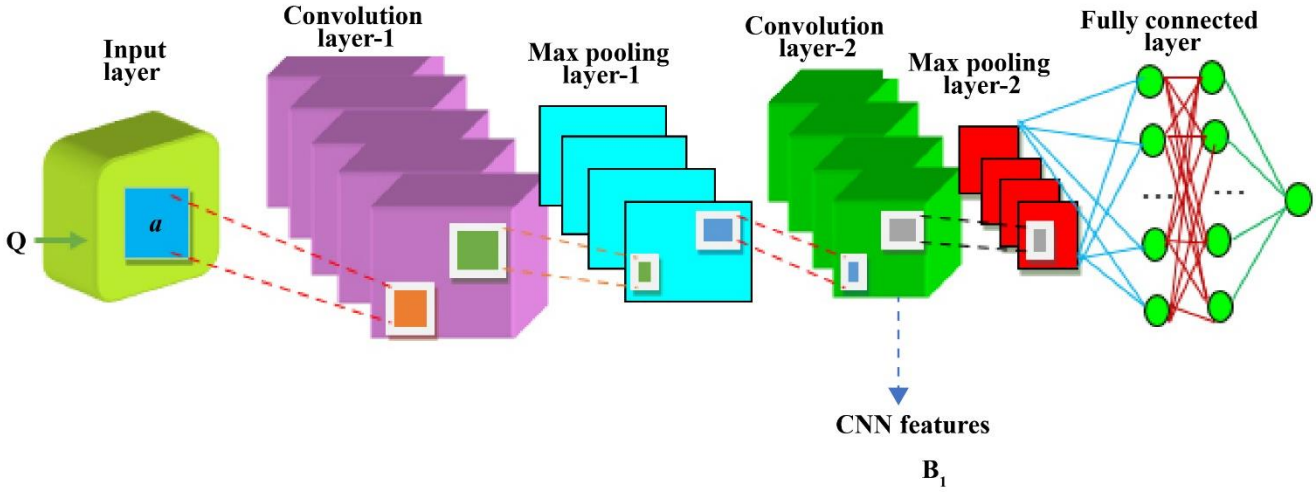


Fig. 3 CNN model

Width

It refers to the width [1] of the lip region, which is denoted as B_6

Location

The location [1] is determined by its "x" and "y" position. The "x" position refers mean of "x"-position pixels in the patch, which is nothing but the central location of the "x"-coordinate. Likewise, the "y"-position signifies the mean y-position of pixels considering each patch, which is nothing but the central location of the "y"-coordinate and is denoted byB_7.

LGTP

LGTP [1] helps to obtain small details of texture and is defiant to lighting changes, and LGTP is the best option for coding the fine details of appearance and texture. In addition, the Gabor features encode shape to the huge range of scale, and its harmonizing nature makes it a good candidate. The LGTP follows four steps. The first is the normalization of the image. The second is convolution with 40 Gabor filters having five scales and eight orientations. The third step is exploiting LGTP for evaluating the generated images, and the fourth is histograms LGTP images and adjoining neighbour classification. It is denoted asB_8.

a) Statistical Features

It includes mean kurtosis, skewness, variance and entropy.

i) Mean

It refers to determining the average pixels present in an image and represented by,

$$\mu = \frac{1}{|e(L_h)|} \times \sum_{h=1}^{|e(L_h)|} e(L_h) \tag{6}$$

where, h refers to overall images, $e(L_h)$signifies the value of a pixel with each image, and $|e(L_h)|$symbolize the total pixel present in pre-processed images. The mean is represented by B_9.

ii) Variance

It is evaluated with the value of mean and is expressed by,

$$\sigma = \frac{\sum_{h=1}^{|e(L_h)|} |L_h - \mu|}{e(L_h)} \tag{7}$$

where, μ indicates mean, $e(L_h)$signifies the value of the pixel using each image, and $L_h$refersh^{th}pixel. The variance feature is denoted byB_{10}.

iii) Kurtosis

Kurtosis $B_{11}$is symmetry and defines the shape of the object.

iv) Skewness

SkewnessB_{12}defines object shape with a numerical value.

v) Entropy

The entropy [29] signifies the metric utilized to determine data uncertainty. In addition, entropy is defined as equivalent intensity states. The entropy is expressed as,

$$Ent = -F\ log(F) \tag{8}$$

where, F refers probability distribution of pixels. The entropy feature is given by B_{13}.

Hence, the feature vector obtained with features extracted from the frame is formulated as,

$$F = \{B_1 B_2, B_3, B_4, B_5, B_6, B_7, B_8, B_9, B_{10}, B_{11}, B_{12}, B_{13}\} \tag{9}$$

where,B_1 refers to CNN features, $B_2$is Gabor features, $B_3$is orientation, B_4represent area, $B_5$is mass, B_6 is width, $B_7$is location, $B_8$is LGTP, and statistical features that include mean, variance, kurtosis, skewness, and entropy are represented asB_9, B_{10},B_{11},$B_{12}$and B_{13}.

3.4.2. Acquirement of Features using Signal

The features attained with signal involve several features wherein each feature is briefly discussed below.

Spectral Centroid

The spectral centroid [25] indicates a frequency-weighted summation of the magnitude spectrum of the normalized signal in contrast to its unweighted sum and is formulated as,

$$B_{14} = \frac{\sum_{\ell=0}^{N-1} \ell|\lambda(\ell)|}{\sum_{\ell=0}^{N-1} |\lambda(\ell)|} \quad (10)$$

where, $\lambda(\ell)$ signifies absolute DFT value, ℓ refers to constant in such a way that$\ell = 0,1,2, \dots, N-1$. The spectral centroid is denoted as B_{14}.

Tonal Power

It is employed to compute the tonalness of the signal [24]. Assume $S_k$refers signal, and $Y(o,f)$ symbolizes voice signal spectrum and is represented as,

$$B_{15} = \frac{X(q)}{\sum_{w=0}^{\frac{f}{2}-1} |Y(o,q)|^2} \quad (11)$$

where, $X(q)$ symbolize tonal power obtained by integrating all bins 0. This tonal power feature is represented byB_{15}.

Spectral Spread

Spectral spread [25] is termed as instant bandwidth that concentrates on magnitude spectrum and is represented as,

$$B_{16} = \frac{\sum_{\ell=0}^{N-1} (\ell - B_{14})^2 |\lambda(\ell)|}{\sum_{\ell=0}^{N-1} |\lambda(\ell)|} \quad (12)$$

where, B_{14}signifies spectral centroid. B_{16} depicts spectral spread.

Pitch Chroma

It provides a basis for devising acoustic patterns. Here, the pitch class indicates a group of pitches that distribute the same chroma. Here, the pitch chroma is expressed as B_{17}.

Spectral Flux

It is used to evaluate how rapidly the signal spectrum is changing and is modelled as,

$$B_{18} = \sum_{\kappa=1}^{Wkl_v} \left(\varpi_{\vartheta}(\kappa) - \varpi_{\vartheta-1}(\kappa)\right)^2 \quad (13)$$

where,$\varpi_{\vartheta}(\kappa)$indicates$\kappa^{th}$ normalized coefficients of DFT at ϑ^{th}frame, $\varpi_{\vartheta-1}(\kappa)$signifies$\kappa^{th}$ normalized coefficient of DFT at $(\vartheta-1)^{th}$frame, $Wk_l$is the total coefficient of DFT. The spectral flux is expressed by B_{18}.

BFCC

The BFCC [22][26] comprises a power spectrum, and its frequencies are converted to bark scale by,

$$B_k(u) = 13\, arctan(0.00076u) + 3.5\, arctan\left(\left(\frac{u}{7500}\right)^2\right) \quad (14)$$

where, B_k indicates bark frequency, and $u$signifies frequency (Hertz). The features mined with BFCC are expressed byB_{19}.

MKMFCC

The MKMFCC [23][27] feature is used to derive multiple kernel-weighted functions. The steps followed to mine MKMFCC are examined as follows:

a) Pre-Emphasis

It is adapted to match modulating power of the signal with a ratio of deviation and is represented by,

$$S(\eta) = Q(\eta) - M * Q(\eta - 1) \quad (15)$$

where, M signifies a value of constant, Qsymbolize input signal, Srefers output signal, and ηrepresent audio signal.

b) Framing

Audio signal instances are split into K blocks of H samples.

c) Hamming Windowing

It is represented by $\varepsilon(\eta): 1 \le \eta \le \hbar - 1$. The signal after performing the windowing is represented as,

$$S(\eta) = Q(\eta) * \varepsilon(\eta) \quad (16)$$

where,$Q(\eta)$ signifies input signal,$\varepsilon(\eta)$ denote hamming window.

d) FFT

The FFT is used to boost discriminating features and is given by,

$$J_v(i) = \frac{1}{l}|L_v(i)|^2 \quad (17)$$

The Discrete Fourier Transform (DFT) of the block is computed as,

$$L_v(i) = \sum_{k=1}^{l} S(\eta).\, e^{-2\pi i i k}; 1 \le i \le o \quad (18)$$

where, i signifies the length of DFT, and$S(\eta)$ covers$l$samples.

e) Mel filter Bank Processing

In processing the Mel filter bank, the signal frequencies are generated with a triangular filter and are formulated by,

$$Mel(\partial) = 1125 \times ln\left(1 + \frac{\partial}{700}\right) \quad (19)$$

where $\partial = 1to \rightleftarrows N$, and symbolize Mel filters number.

f) Filter Bank Energy

The filter bank energy is represented as,

$$\in (b) = \sum log \mid B(h)|D(h)\left(d\frac{2\pi}{J}\right) \times \omega_h \quad (20)$$

where, $\omega_h$symbolize multi-kernel weighted function, and $B(h)$ and $D(h)$ are power spectrums, and $J$is coefficients and d are constant.

g) DCT

In DCT, the log Mel spectrum transformation computes the spatial domain to perform the transformation.

h) Delta Energy and Spectrum

The structures of produced energies are combined using an acoustic feature vector.

i) Cepstral Normalization

Here, the coefficient average is subtracted and divided through variance. Hence, the mined MKMFCC feature is expressed byB_{20}.

Hence, the feature vector obtained with features extracted from the signal is formulated as,

$$D = \{B_{14}, B_{15}, B_{16}, B_{17}, B_{18}, B_{19}, B_{20}\} \quad (21)$$

where, $B_{14}$is the spectral centroid, B_{15}represent Tonal power ratio, B_{16}indicate Spectral spread, B_{17}refers to Pitch chroma, B_{18}indicates Spectral flux, B_{19}signifies BFCC and B_{20}express MKMFCC.

3.4.3. Feature Concatenation

The feature concatenation combines the signal features and mining through the lip region. The combined feature is represented as$E$, which is formulated as,

$$E = B + D \quad (22)$$

where B is features are extracted from the lip region, and Dexpress features are extracted from the signal.

3.5. Visual Speech Recognition using Proposed SEOA-based DRN

The recognition of visual speech signals was performed with SEOA-based DRN. The feature vector$E$ is employed in DRN for visual speech recognition. The DRN training is performed with SEOA and is devised by unifying SOA [38] and EWMA [9]. The DRN model and training with SEOA are examined.

3.5.1. Architecture of DRN

Here, DRN [39] is used for the improved decision in which the decision regarding speech recognition is performed. Figure 4 shows the DRN model.

Convolutional (conv) Layer

The 2D conv layer reduces free parameters in the training phase. The conv layer is computed by,

$$\Re(Q) = \sum_{v=0}^{\gamma-1}\sum_{\chi=0}^{\gamma-1} X_{a,s} \bullet Q_{(u+a),(v+s)} \quad (23)$$

$$\Re(Q) = \sum_{Z=0}^{C_{in}-1} G_Z * Q \quad (24)$$

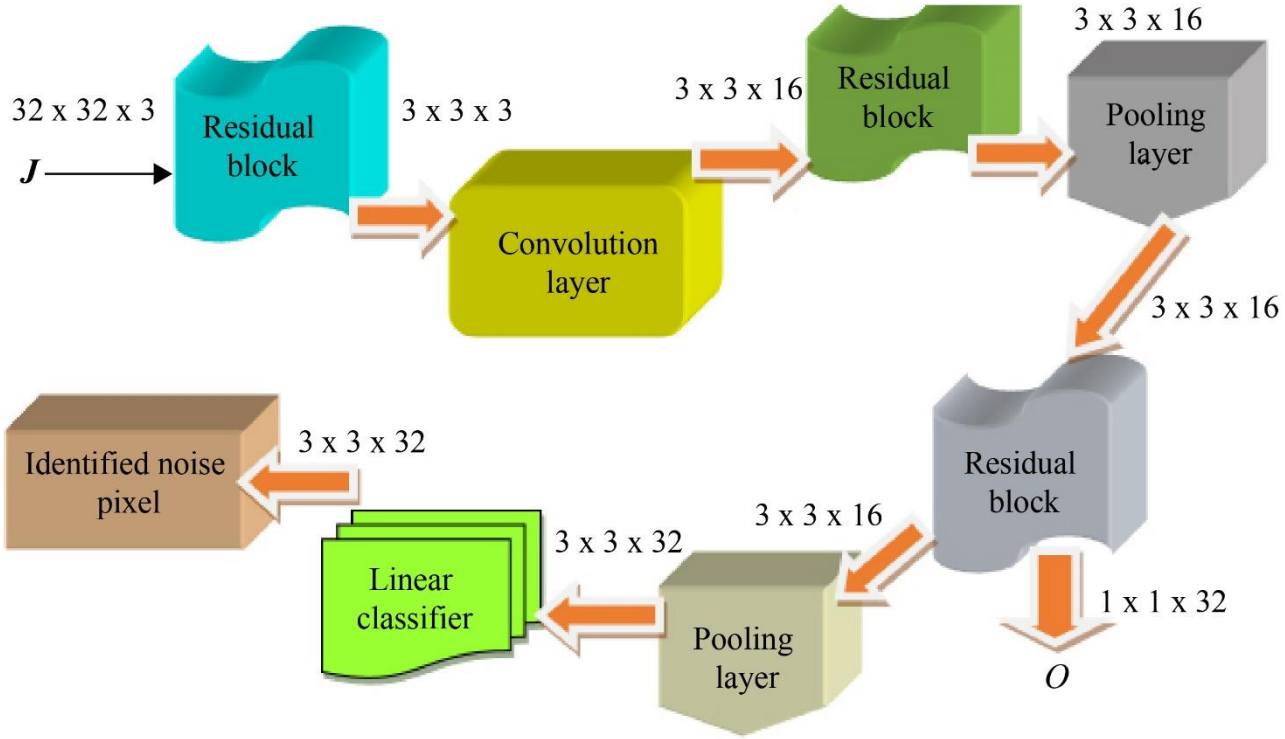


Fig. 4 DRN model

where, Q signifies the CNN feature of the input image, $u$and $v$is used for recording coordinates, $G$is$\gamma \times \gamma$kernel matrix, $v$and$\chi$are kernel matrix index. Hence, $G_Z$represent the size of the kernel in Z^{th} neuron, and $*$ is a cross-correlation operator.

Pooling Layer

This layer is associated with the conv layer and is principally utilized to reduce spatial size. Hence, average pooling is chosen to function and is formulated by.

$$r_{out} = \frac{r_{in} - \kappa_a}{\lambda} + 1 \quad (25)$$

$$i_{out} = \frac{i_{in} - \kappa_s}{\lambda} + 1 \quad (26)$$

where, r_{in} refers width of the input matrix, i_{in} symbolize input matrix height, $r_{out}$and i_{out}denote the value of individual output. Furthermore, κ_a refers width and κ_s deliberates the height of kernel size.

Activation Function

The non-linear activation function is employed for learning features and is given by,

$$Re\,L\,U(Q) = \begin{cases} 0; \ell < 0 \\ \ell; \ell \geq 0 \end{cases} \quad (27)$$

Here, ℓrefers feature.

Batch Normalization

The training set is split into several tiny sets termed mini-batches to train the model.

Residual Blocks

It refers to shortcut association among the conv layer and is given by,

$$O_l = \Re(Q) + Q \quad (28)$$

$$O = \Re(Q) + \lambda_M Q \quad (29)$$

Here, Q signify input and $O_l$are output residual blocks, Osignifies mapping relations, and λ_M express dimension matching factor.

Linear Classifier

After evaluating the conv layer, the linear classifier adopts a process to determine noisy pixels with the input image and is given by,

$$O = \lambda O + v \quad (30)$$

Here, λ is the weight matrix and $v$symbolize bias. The obtained output is denoted as Othat assists in visual speech recognition.

3.5.2. DRN Training with SEOA

The weight update of DRN is done using SEOA and obtained by blending SOA and EWMA. SOA [38] is an algorithm that mimics the social features of human beings in society. It performs effectively on engineering design issues. Meanwhile, EWMA [9] is developed by averaging a number of successive observations in which values are weighed. It is adapted to identify tiny shifts in handing out targeted value. Hence, the combination of EWMA and SOA improves complete performance and assists in providing an enhanced solution.

Step 1) Initialization

The first step is the initialization of the solution and given by,

$$G = \{G_1, G_2, \ldots, G_\rho, \ldots, G_\kappa\} \quad (31)$$

where, κ symbolizes total solution, and $G_\rho$express the ρ^{th}solution.

Step 2) Find an Error

The most advantageous solution is obtained with MSE and is given by,

$$MS_{err} = \frac{1}{g}\sum_{h=1}^{g}[\xi_h - O]^2 \quad (32)$$

where, ξ_h is expected output, and Odeliberates output produced with DRN, $g$refers data count.

Step 3) Enumerate Equality of Opportunity

According to SOA [38], the equality of opportunity depicts a fairness-seeking principle, and it reveals the resulting position of each entity and is expressed as,

$$G_l(q+1) = G_l(q) + rand(P - Q \times G_l(q)) \quad (33)$$

Where, $r\ and$ expresses arbitrary number, Pindicates best position, Qsymbolize personal choice coefficient, $G_l(q)$ represent old entity location. The best solution is expressed as,

$$P = rand\{S, A\} \quad (34)$$

where S is the best solution, and$A$symbolize density point. The density point is expressed by,

$$A = \frac{D_1 z_1 + D_2 z_2 + \cdots + D_o z_o}{z_1 + z_2 + \cdots + z_o} \quad (35)$$

where, $z_1$refers to results generated by a society member.

As per EWMA [9], the update is provided by,

$$G_l^E(q) = \sigma G_l(q) + (1 - \sigma) * G_l^E(q-1) \quad (36)$$

where, $\sigma$is exponential constant, and $G_l(q)$ refers current solution and $G_l^E(q-1)$ express prior solution.

$$G_l(q) = \frac{G_l^E(q) - (1-\sigma) * G_l^E(q-1)}{\sigma} \quad (37)$$

Substitute equation (37) in equation (33),

$$G_l(q+1) = \frac{G_l^E(q)-(1-\sigma)*G_l^E(q-1)}{\sigma} + rand\left(P - Q \times \frac{G_l^E(q)-(1-\sigma)*G_l^E(q-1)}{\sigma}\right) \quad (38)$$

$$G_l(q+1) = \frac{G_l^E(q)}{\sigma} - \frac{(1-\sigma)*G_l^E(q-1)}{\sigma} + rand\left(P - Q \times \frac{G_l^E(q)-(1-\sigma)*G_l^E(q-1)}{\sigma}\right) \quad (39)$$

$$G_l(q+1) = \frac{G_l^E(q)}{\sigma} - \frac{(1-\sigma)*G_l^E(q-1)}{\sigma} + randP - randQ \times \frac{G_l^E(q)}{\sigma} + randQ \times \frac{(1-\sigma)*G_l^E(q-1)}{\sigma} \quad (40)$$

$$G_l(q+1) = \frac{G_l^E(q)}{\sigma}(1-randQ) - \frac{(1-\sigma)*G_l^E(q-1)}{\sigma}(1-randQ) + randP \quad (41)$$

Step 4) Enumerate the principle of community
It is given as,

$$G_l(q+1) = G_l(q) + rand(S-J) \quad (42)$$

Substitute equation (37) in equation (42),

$$G_l(q+1) = \frac{G_l^E(q)-(1-\sigma)*G_l^E(q-1)}{\sigma} + rand(S-J) \quad (43)$$

where, Jrefers to the empty point, and S refers to the best solution. The empty point is formulated by,

$$J = \frac{D_1\frac{1}{z_1}+D_2\frac{1}{z_2}+\cdots+D_o\frac{1}{z_o}}{\frac{1}{z_1}+\frac{1}{z_2}+\cdots+\frac{1}{z_o}} \quad (44)$$

Step 5) Enumerate the empty point and density point
The density point is expressed by,

$$A = \sum_{l=1}^{I} \frac{z_l}{\sum_{r=1}^{I} z_r} D_l \quad (45)$$

The empty point is expressed by,

$$J = \sum_{l=1}^{I} \frac{\frac{1}{z_l}}{\sum_{r=1}^{I}\frac{1}{z_l}} D_l \quad (46)$$

where, $\frac{z_l}{\sum_{r=1}^{I} z_r}$ is relative fitness. The density is expressed by,

$$A = \sum_{l=1}^{I} \hat{z}_l D_l \quad (47)$$

The empty points are expressed by,

$$J = \sum_{l=1}^{I} \check{z}_l D_l \quad (48)$$

where, z_l is outcomes generated by a society member.
Here, $\hat{z}$is modelled by,

$$\hat{z}_l = \frac{e^{\frac{z_l}{z_{max}}}}{\sum_{r=1}^{I\Sigma} e^{\frac{z_r}{z_{max}}}} \quad (49)$$

where, $z_r$is the result generated by society members.
Here, $\check{z}$is modelled by,

$$\check{z}_l = \frac{e^{-\frac{z_l}{z_{max}}}}{\sum_{r=1}^{I\Sigma} e^{-\frac{z_r}{z_{max}}}} \quad (50)$$

Step 6) Re-evaluate error for update solutions
The error of update solutions is re-calculated in which the optimum solution is obtained.

Step 7) Terminate
The optimum solutions are derived iteratively. The pseudo-code of SEOA is provided in table 1.

Table 1. Pseudo code of SEOA

Input: $\boldsymbol{G}$: Solutions Set, $\boldsymbol{q}$: present iteration, $\boldsymbol{q_{max}}$: highest iteration
Output: Optimum solution $\boldsymbol{G}^*$
Begin
Initialize population $\boldsymbol{G}$ randomly;
Enumerate population;
$\boldsymbol{S}$ ←Best solution;
For $\boldsymbol{q = 1 to q_{max}}$do
$\boldsymbol{J}$ ← Enumerate empty point;
For $\boldsymbol{l = 1 to I}$do
$\boldsymbol{E = rand(S,J)}$;
$\boldsymbol{F = rand\{0,1,2\}}$;
Enumerate $\boldsymbol{G_l^{new}}$with equation (43)
end
If$\boldsymbol{G_l(q+1)}$ better than $\boldsymbol{G_l(q)}$ then
$\boldsymbol{G_l(q) \leftarrow G_l(q+1)}$
end
Enumerate new population
$\boldsymbol{S}$ ←Best solution;
$\boldsymbol{A}$ ← Enumerate Density point;
For$\boldsymbol{q = 1 to q_{max}}$do
Enumerate $\boldsymbol{G_l(q+1)}$ with equation (41)
end
If$\boldsymbol{G_l(q+1)}$ better than $\boldsymbol{G_l(q)}$ then
$\boldsymbol{G_l(q) \leftarrow G_l(q+1)}$
end
Enumerate new population
$\boldsymbol{S}$ ←Best solution;
end
Acquire the best solution

The output of the proposed SEOA-based DRN is expressed as Othat helps in visual speech recognition.

4. Results and Discussions

The proficiency of the proposed SEOA-based DRN is calculated by changing training data and K-fold. Furthermore, the SEOA-based DRN is evaluated by varying the population size.

4.1. Experimental Setup

The SEOA-based DRN is executed on Windows 10 OS with 8GB RAM and Intel core i5 processor and executed in MATLAB.

4.2. Dataset Used

The analysis is performed using Oxford-BBC Lip Reading in the Wild (LRW) Dataset [18]. This dataset comprises 1000 utterances of 500 different words, which various speakers articulate. The video contains 29 frames, and words occur in the middle of the video. The duration of the word is provided in metadata that can be discovered from the start and end frames.

4.3. Experimental Outcomes

The upshots are taken with a set of video frames and voice signals briefly described below.

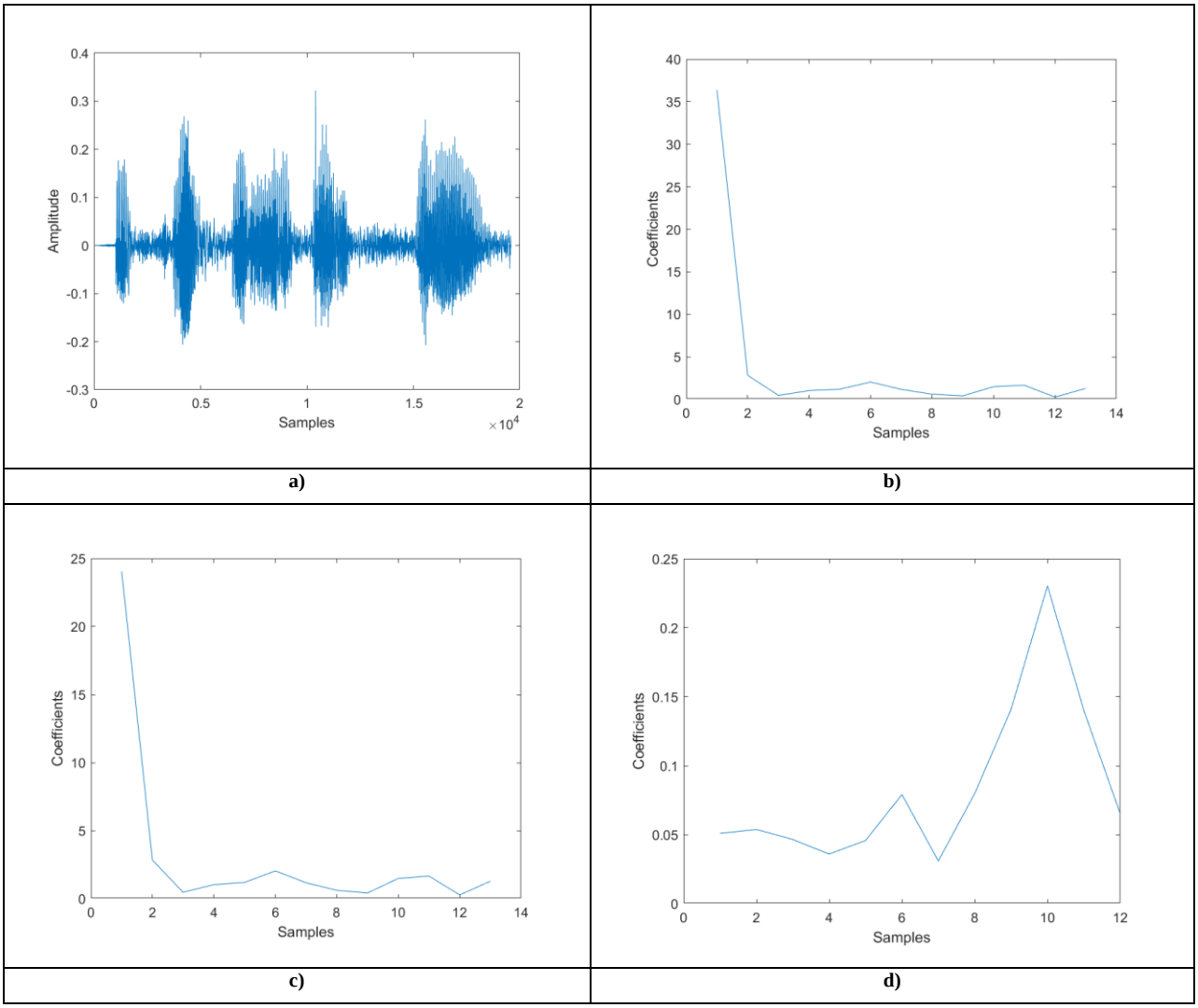


Fig. 5 Experimental outcomes of SEOA-based DRN with a) Input voice sample

b) BFCC extracted signal c) MKMFCC extracted signal d) pitch chroma extracted signal

4.3.1. Using a Voice Signal

Figure 5 shows the experimental outcomes of SEOA-based DRN considering the set of the input voice sample. The inputted voice samples taken for the analysis are exposed in figure 5a). The BFCC extracted signal is shown in figure 5b). The MKMFCC mined signal is shown in figure 5c). The pitch chroma extracted signal is shown in figure 5d).

4.3.2. Using Video Frame

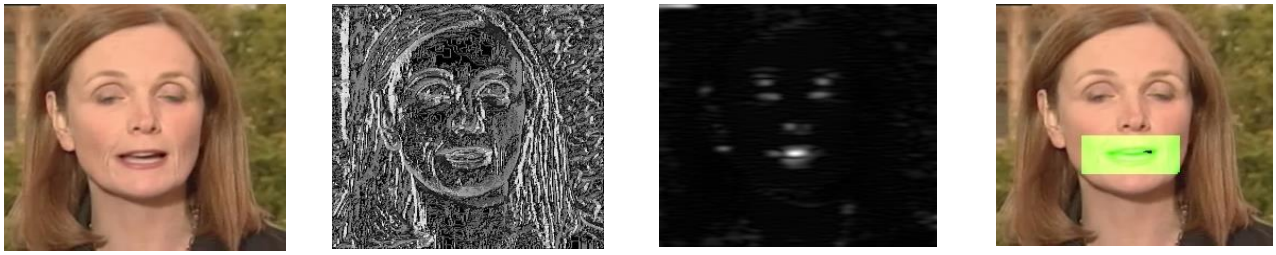
Figure 6 exposes the experimental outcomes of SEOA-based DRN with video frames. The inputted video frames are shown in figure 6a). The LGTP feature extracted frame is shown in figure 6b). The Gabor features mined frame is shown in figure 6c). The lip segmentation frame is shown in figure 6d.)

4.4. Evaluation Measures

The adaption of the developed SEOA-based DRN is done with the following chosen metrics

4.4.1. Accuracy

It indicates the nearness degree of calculated value in contrast to the original value in visual speech recognition.



a) b) c) d)

Fig. 6 Experimental outcomes of SEOA-based DRN considering a) Input video frame b) LGTP feature extracted frame c) Gabor feature extracted frame d) Lip segmented frame

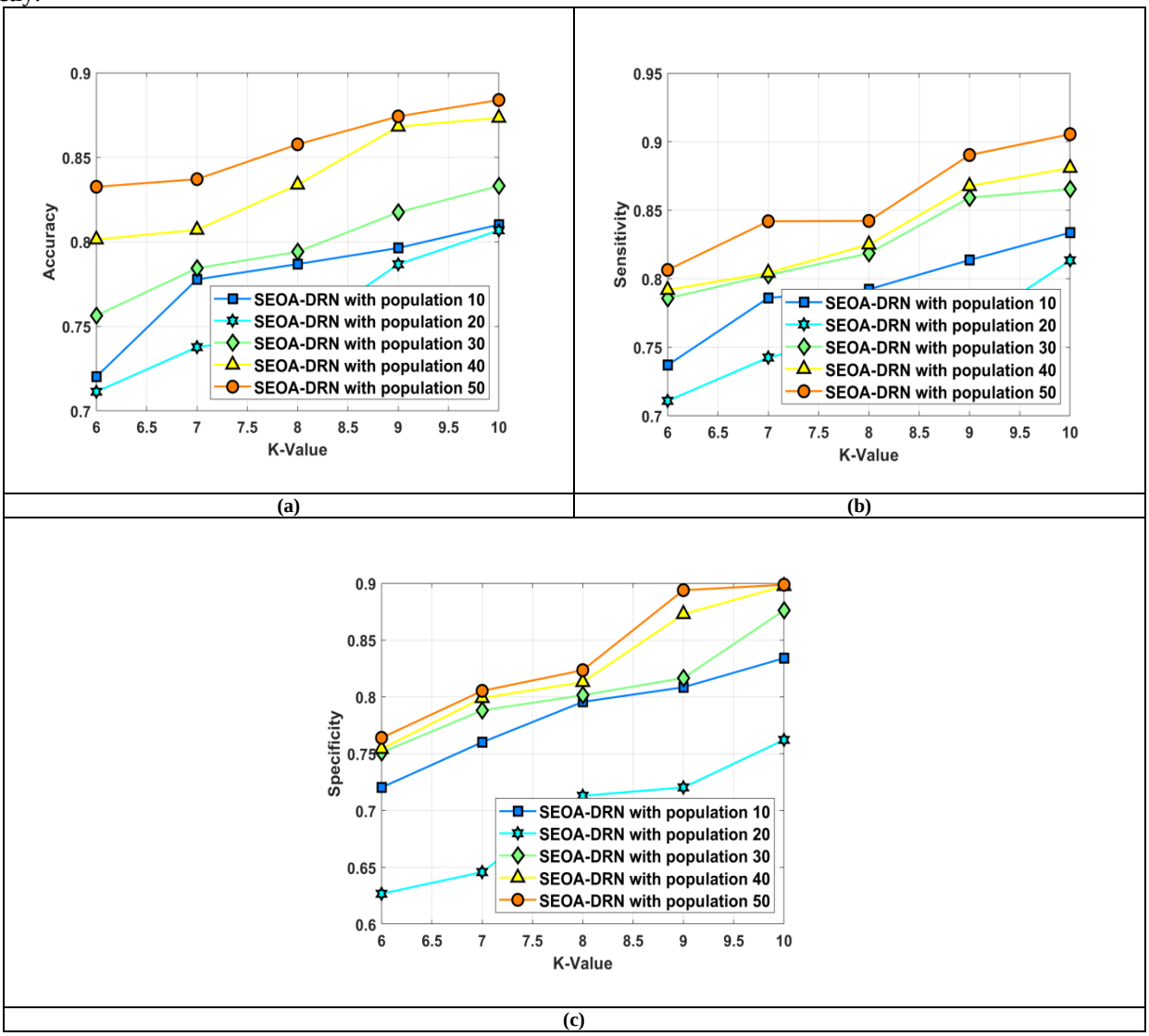
4.5. Performance Analysis

4.5.1. Sensitivity

It refers to the proportion of positives and is determined by the visual speech recognition method exactly.

4.5.2. Specificity

It refers to the proportion of negatives discovered by the developed model specifically.



(a) (b)

(c)

Fig. 7 Performance assessment of SEOA-DRN considering a) Accuracy, b) Sensitivity, c) Specificity

Figure 7 exposes the performance assessment of SEOA-DRN by varying K-values. The accuracy assessment is depicted in figure 7a). For K-value=6, the accuracy evaluated by SEOA-DRN with populations 10, 20, 30, 40, and 50 is 0.720, 0.711, 0.756, 0.801, and 0.833. Similarly, for K-value=10, the accuracy evaluated by SEOA-DRN with populations 10, 20, 30, 40, and 50 are 0.810, 0.807, 0.833, 0.873, and 0.884. The sensitivity assessment is revealed in figure 7b). For K-value=6, the sensitivity evaluated by SEOA-DRN with populations 10, 20, 30, 40, and 50 is 0.737, 0.711, 0.786, 0.792, and 0.806. Similarly, for K-value=10, the sensitivity evaluated by

SEOA-DRN with populations 10, 20, 30, 40, and 50 are 0.834, 0.813, 0.865, 0.881, and 0.906. The specificity assessment is revealed in figure 7c). For K-value=6, the specificity evaluated by SEOA-DRN with populations 10, 20, 30, 40, and 50 is 0.720, 0.627, 0.751, 0.754, and 0.764. Similarly, for K-value=10, the specificity evaluated by SEOA-DRN with--population 10, 20, 30, 40, and 50 are 0.834, 0.762, 0.876, 0.898, and 0.899.

4.6. Comparative Methods

The techniques taken for assessment include BiLSTM [1], CNN [7], CNN+BiLSTM [8], ResNet [3] and proposed SEOA-DRN.

4.7. Comparative Analysis

The assessment considering specificity, accuracy and sensitivity are described by altering training data and K-fold.

Figure 8 displays the assessment of techniques by varying K-values. The accuracy assessment is depicted in figure 8a). For 50% of training data, the accuracy enumerated by BiLSTM, CNN, CNN+BiLSTM, ResNet and proposed SEOA-DRN is 0.720, 0.711, 0.756, 0.801, and 0.833. Similarly, for 90% of training data, the accuracy was enumerated.

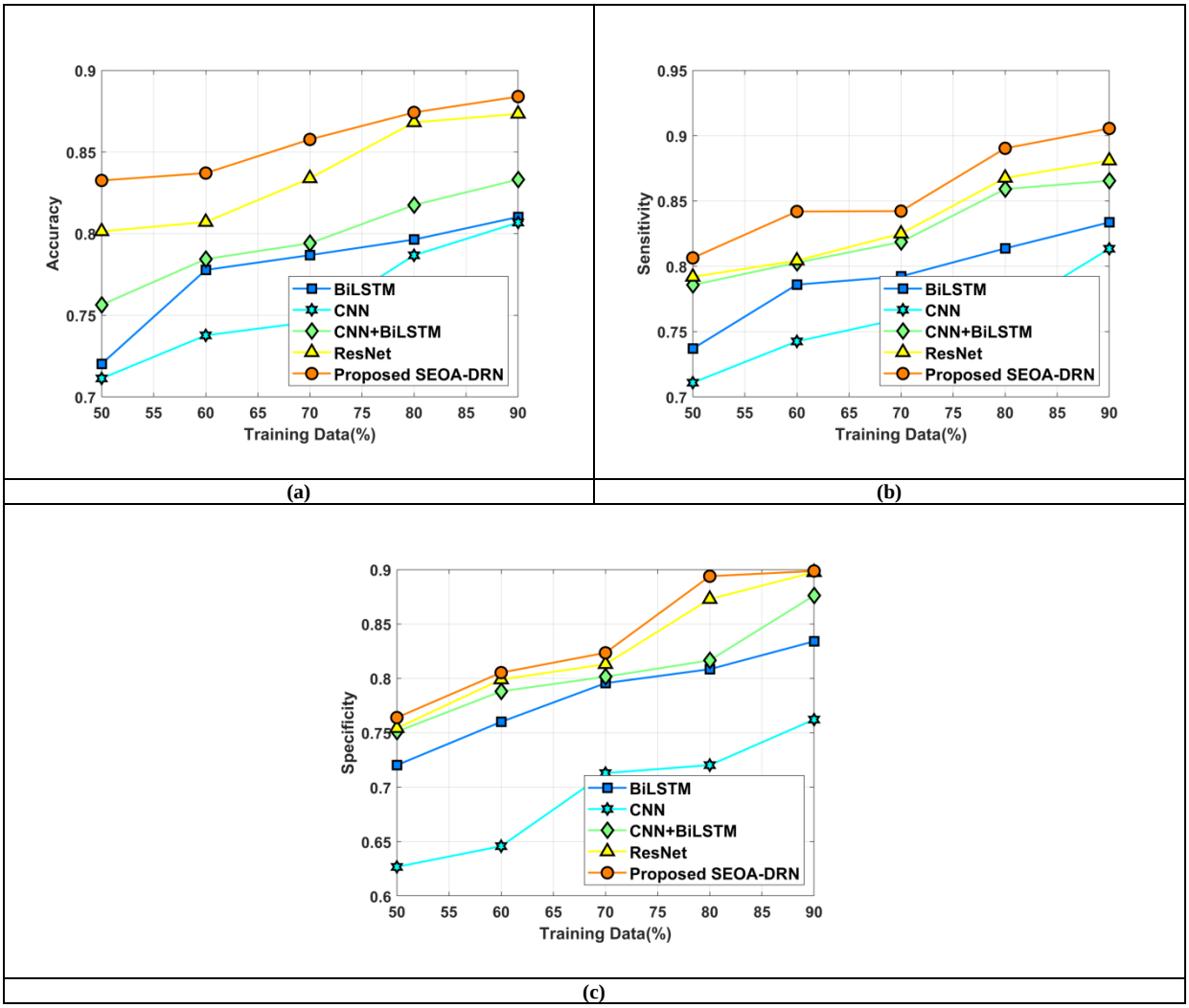


Fig. 8 Assessment by changing training data considering a) Accuracy, b) Sensitivity, c) Specificity

The sensitivity assessment is depicted in figure 8b). For 50% of training data, the sensitivity enumerated by BiLSTM, CNN, CNN+BiLSTM, ResNet and proposed SEOA-DRN is 0.737, 0.711, 0.786, 0.792, and 0.806. Similarly, for 90% of training data, the sensitivity enumerated by BiLSTM, CNN, CNN+BiLSTM, ResNet and proposed SEOA-DRN is 0.834, 0.813, 0.865, 0.881, and 0.906. The efficiency of BiLSTM, CNN, CNN+BiLSTM, and ResNet in contrast to the proposed SEOA-DRN using sensitivity are 7.947%, 10.264%, 4.525%, and 2.814%. The specificity assessment is depicted in figure 8c). For 50% of training data, the specificity enumerated by BiLSTM, CNN, CNN+BiLSTM, ResNet and proposed SEOA-DRN is 0.720, 0.627, 0.751, 0.754, and 0.764. Similarly, for 90% of training data, the specificity enumerated by BiLSTM, CNN, CNN+BiLSTM, ResNet and proposed SEOA-DRN is 0.834, 0.762, 0.876, 0.898, and 0.899. The efficiency of BiLSTM, CNN, CNN+BiLSTM, and ResNet in contrast to the proposed SEOA-DRN using specificity are 7.230%, 15.239%, 2.558%, and 0.111%.

Figure 9 displays the assessment of techniques by varying K-values. The assessment with accuracy is revealed in figure 9a). For K-value=6, the accuracy enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.741, 0.724, 0.753, 0.759, whereas for proposed SEOA-DRN is 0.793. Similarly, for K-value=10, the accuracy enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.857, 0.836, 0.883, 0.883, whereas for proposed SEOA-DRN is 0.904. The efficiency of BiLSTM, CNN, CNN+BiLSTM, and ResNet in contrast to SEOA-DRN using accuracy are 5.199%, 7.522%, 2.323%, and 2.323%. The assessment with sensitivity is revealed in figure 9b). For K-value=6, the sensitivity enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.765, 0.742, 0.715, 0.788, whereas for proposed SEOA-DRN is 0.823. Similarly, for K-value=10, the sensitivity enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.850, 0.847, 0.835, 0.857, whereas for proposed SEOA-DRN is 0.894. The efficiency of BiLSTM, CNN, CNN+BiLSTM, and ResNet in contrast to SEOA-DRN using sensitivity are 4.921%, 5.257%, 6.599%, and 4.138%. The assessment with specificity is revealed in figure 9c). For K-value=6, the specificity enumerated by BiLSTM, CNN, CNN, CNN+BiLSTM, and ResNet is 0.769, 0.756, 0.751, 0.778, whereas for proposed SEOA-DRN is 0.817. Similarly, for K-value=10, the specificity enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.871, 0.859, 0.831, 0.871, whereas for proposed SEOA-DRN is 0.897. The efficiency of BiLSTM, CNN, CNN+BiLSTM, and ResNet in contrast to SEOA-DRN using specificity are 2.898%, 4.236%, 7.357%, and 2.898%.

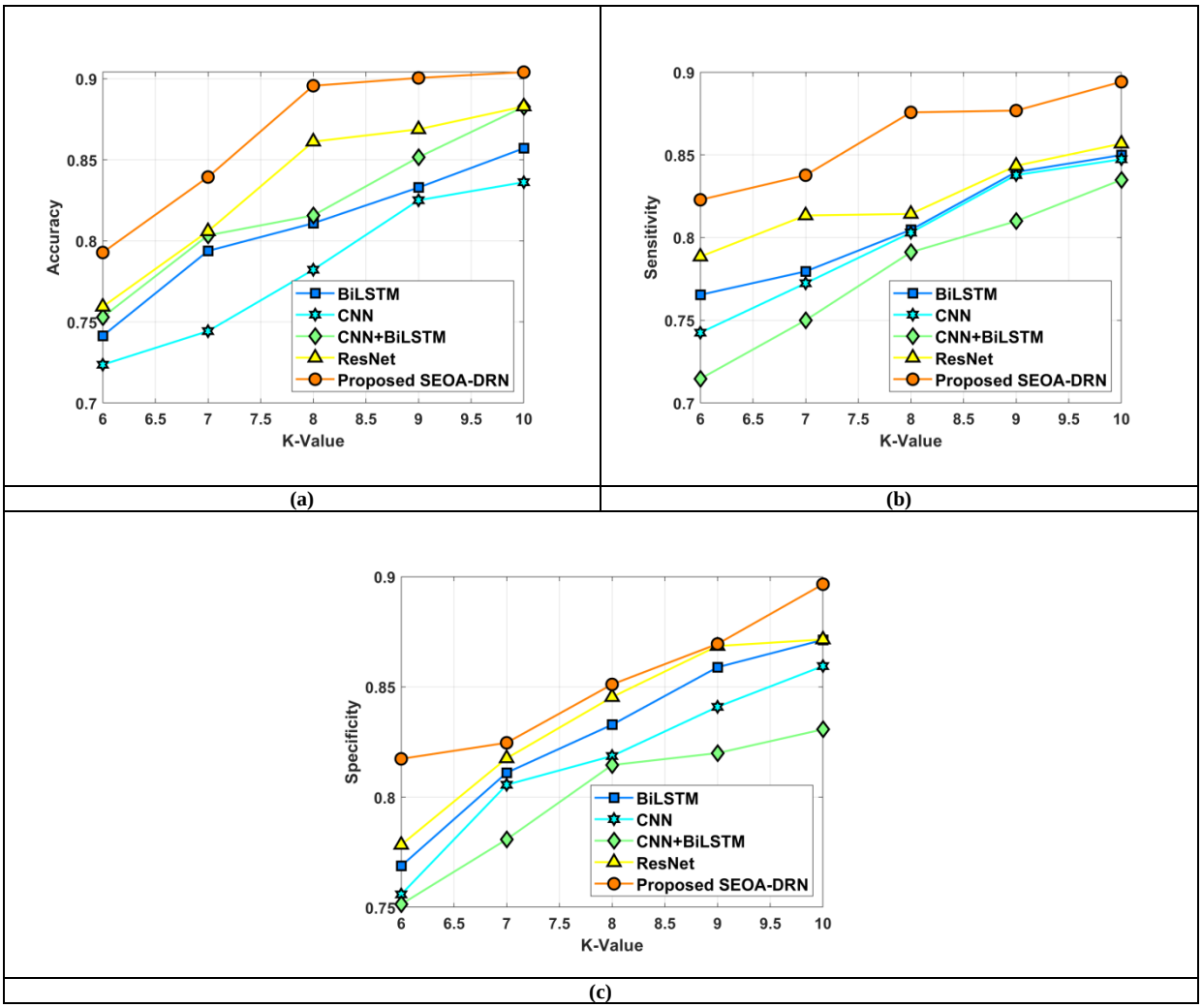


Fig. 9 Assessment by changing K-fold using a) Accuracy, b) Sensitivity, c) Specificity

4.8. Comparative Discussion

Table 2 presents the competence of strategies considering classical models. With training data, the elevated accuracy of 0.884 is enumerated by SEOA-DRN, while the accuracy enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.810, 0.807, 0.833, and 0.873. The highest sensitivity of 0.906 is enumerated by SEOA-DRN, while the sensitivity enumerated by BiLSTM, CNN, CNN+ enumerated, and ResNet is 0.834, 0.813, 0.865, and 0.881. SEOA-DRN measures the highest specificity of 0.899, while the specificity enumerated by BiLSTM, CNN, CNN+BiLSTM, and ResNet are 0.834, 0.762, 0.876, and 0.898. With K-value, the highest accuracy of 0.904, highest sensitivity of 0.894 and highest specificity of 0.897 is measured by SEOA-DRN.

Table 2. Comparative analysis

Variation	Metrics	CNN	BiLSTM	CNN+BiLSTM	ResNet	Proposed SEOA-DRN
Training data	**Accuracy**	0.807	0.810	0.833	0.873	**0.884**
	Sensitivity	0.813	0.834	0.865	0.881	**0.906**
	Specificity	0.762	0.834	0.876	0.898	**0.899**
K-fold	**Accuracy**	0.836	0.857	0.883	0.883	**0.904**
	Sensitivity	0.847	0.850	0.835	0.857	**0.894**
	Specificity	0.859	0.871	0.831	0.871	**0.897**

5. Conclusion

A fresh optimization-driven deep model is developed for lip reading-based visual speech recognition. The goal is to devise an effective lip reading-based visual speech recognition using both audio and video data. The speech recognition system combines visual data features and audio signal features for reliable speech recognition. This technique performed robust extraction of lip region using the CFPNet-M. Thus, lip reading helps the person comprehend the speech by watching and discovering the movements of the mouth linked with speech. The speech information is modelled in some intensity and shape, and several visual and signal features are concatenated to enhance the performance. This method has led to robust recognition of speech using the newly devised model, namely SEOA-DRN, wherein the DRN is employed, and its weights are tuned with SEOA. It employed both visual and signal features to increase recognition performance in noisy platforms. The SEOA-DRN surpass eminent accuracy of 88.4%, the highest sensitivity of 90.6% and the highest specificity of 90.6%. The future work involves deliberation of other databases to increase the feasibility of the designed strategy.

References

[1] Xuejie Zhang et al., "Visual Speech Recognition with Lightweight Psychologically Motivated Gabor Features," *Entropy*, vol. 22, no. 12, p. 1367, 2020. *Crossref*, https://doi.org/10.3390/e22121367

[2] Stavros Petridis et al., "End-to-End Visual Speech Recognition for Small-Scale Datasets," *Pattern Recognition Letters*, vol. 131, pp. 421-427, 2020. *Crossref*, https://doi.org/10.1016/j.patrec.2020.01.022

[3] H Liu, Z Chen, and B Yang, "Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion," *Proceedings of INTERSPEECH*, pp. 3520-3524, 2020. *Crossref*, https://doi.org/10.21437/Interspeech.2020-3146

[4] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-To-End Audio-Visual Speech Recognition with Conformers," *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7613-7617, 2021. *Crossref*, https://doi.org/10.1109/ICASSP39728.2021.9414567

[5] Nilay Shrivastava et al., "MobiVSR: Efficient and Light-Weight Neural Network for Visual Speech Recognition on Mobile Devices," *Proceedings of INTERSPEECH*, pp. 2753-2757, 2019. *Crossref*, https://doi.org/10.21437/Interspeech.2019-3273

[6] Abitha A, and Lincy K, "A Faster RCNN Based Image Text Detection and Text to Speech Conversion," *SSRG International Journal of Electronics and Communication Engineering*, vol. 5, no. 5, pp. 11-14, 2018. *Crossref*, https://doi.org/10.14445/23488549/IJECE-V5I5P103

[7] T. Ozcan, and A. Basturk, "Lip Reading using Convolutional Neural Networks with and without Pre-Trained Models," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 2, pp. 195-201, 2019. *Crossref*, https://doi.org/10.17694/bajece.479891

[8] Yuanyao Lu, and Jie Yan, "Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-Term Memory," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 101, p. 2054003, 2020. *Crossref*, https://doi.org/10.1142/S0218001420540038

[9] Michael S. Saccucci, Raid W. Amin, and James M. Lucas, "Exponentially Weighted Moving Average Control Schemes with Variable Sampling Intervals," *Communications in Statistics-Simulation and Computation*, vol. 21, no.3, pp. 627-657, 1992. *Crossref*, https://doi.org/10.1080/03610919208813040

[10] Ayush Jain et al., "Detection of Sarcasm through Tone Analysis on video and Audio files: A Comparative Study on AI Models Performance," *SSRG International Journal of Computer Science and Engineering*, vol. 8, no. 12, pp. 1-5, 2021. *Crossref*, https://doi.org/10.14445/23488387/IJCSE-V8I12P101

[11] Adriana Fernandez-Lopez, and Federico M.Sukno, "Survey on Automatic Lip-Reading in the Era of Deep Learning," *Image and Vision Computing*, vol. 78, pp. 53-72, 2018. *Crossref*, https://doi.org/10.1016/j.imavis.2018.07.002

[12] George Sterpu, and Naomi Harte, "Towards Lipreading Sentences with Active Appearance Models," 2018.

[13] Nancy Tye-Murray et al., "Lipreading in School-Age Children: The Roles of Age, Hearing Status, and Cognitive Ability," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 2, pp. 556-565, 2014. *Crossref*, https://doi.org/10.1044/2013_JSLHR-H-12-0273

[14] C. Senthil Kumar, and Y. Raj Kumar, "Bus Embarking System for Visual Impaired People using Radio-Frequency Identification," *SSRG International Journal of Electronics and Communication Engineering*, vol. 4, no. 4, pp. 10-15, 2017. *Crossref*, https://doi.org/10.14445/23488549/IJECE-V4I4P103

[15] Yiting Li et al., "Lip Reading Using a Dynamic Feature of Lip Images and Convolutional Neural Networks," *Proceedings of 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1-6, 2016. *Crossref*, https://doi.org/10.1109/ICIS.2016.7550888

[16] Timothy Israel Santos, and Andrew Abel, "Using Feature Visualisation for Explaining Deep Learning Models in Visual Speech," *Proceedings of 2019 IEEE 4th International Conference on Big Data Analytics*, pp. 231-235, 2019. *Crossref,* https://doi.org/10.1109/ICBDA.2019.8713256
[17] Andrew Abel et al., "A Data Driven Approach to Audiovisual Speech Mapping," *Proceedings of International Conference on Brain Inspired Cognitive Systems*, pp. 331-342, 2016. *Crossref,* https://doi.org/10.1007/978-3-319-49685-6_30
[18] The Oxford-BBC Lip Reading in the Wild (LRW). [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html
[19] Ravi Kumar Saini, and Mrs. Mamta Yadav, "Image & Video Quality Assessment and Human Visual Perception," *International Journal of Computer & organization Trends*, vol. 6, no. 3, pp. 1-4, 2016. *Crossref,* https://doi.org/10.14445/22492593/IJCOT-V34P301
[20] Ange Lou, Shuyue Guan, and Murray Loew, "CFPNet-M: A Light-Weight Encoder-Decoder Based Network for Multimodal Biomedical Image Real-Time Segmentation," 2021.
[21] Wai Chee Yau, Hans Weghorn, and Dinesh Kant Kumar, "Visual Speech Recognition and Utterance Segmentation Based on Mouth Movement," *Proceedings of 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications*, pp. 7-14, 2007. *Crossref,* https://doi.org/10.1109/DICTA.2007.4426769
[22] Chandar Kumar et al., "Analysis of MFCC and BFCC in a Speaker Identification System," *Proceedings 2018 International Conference on Computing, Mathematics and Engineering Technologies*, pp. 1-5, 2018. *Crossref,* https://doi.org/10.1109/ICOMET.2018.8346330
[23] Osama S. Faragallah, "Robust Noise MKMFCC–SVM Automatic Speaker Identification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 185-192, 2018. *Crossref,* https://doi.org/10.1007/s10772-018-9494-9
[24] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman, "A Novel Adaptive Fractional Deep Belief Networks for Speaker Emotion Recognition," *Alexandria Engineering Journal*, vol. 56, no. 4, pp. 485-497, 2017. *Crossref,* https://doi.org/10.1016/j.aej.2016.09.002
[25] Ahnaf Rashik Hassan, and Mohammad Aynal Haque, "Computer-Aided Sleep Apnea Diagnosis from Single-Lead Electrocardiogram Using Dual Tree Complex Wavelet Transform and Spectral Features," *Proceedings of International Conference on Electrical & Electronic Engineering*, pp. 49-52, 2015. *Crossref,* https://doi.org/10.1109/CEEE.2015.7428289
[26] Lichuan Liu et al., "Infant Cry Language Analysis and Recognition: An Experimental Approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 3, pp. 778-788, 2019. *Crossref,* https://doi.org/10.1109/JAS.2019.1911435
[27] Arul Valiyavalappil Haridas et al., "Emotion Recognition of Speech Signal Using Taylor Series and Deep Belief Network Based Classification," *Evolutionary Intelligence*, pp. 1145-1158, 2022. *Crossref,* https://doi.org/10.1007/s12065-019-00333-3
[28] Siti Nurulain Mohd Rum, and Braxton Isaiah Boilis, "Sign Language Communication through Augmented Reality and Speech Recognition (LEARNSIGN)," *International Journal of Engineering Trends and Technology*, vol. 69, no. 4, pp. 125-130, 2021. *Crossref,* https://doi.org/10.14445/22315381/IJETT-V69I4P218
[29] Pei Wang, Hui Fu, and Ke Zhang, "A Pixel-Level Entropy-Weighted Image Fusion Algorithm Based on Bidimensional Ensemble Empirical Mode Decomposition," *International Journal of Distributed Sensor Networks*, vol. 14, no. 12, 2018. *Crossref,* https://doi.org/10.1177/1550147718818755
[30] J.S. Anita, and J.S. Abinaya, "Impact of Supervised Classifier on Speech Emotion Recognition," *Multimedia Research*, vol. 2, no. 1, pp. 9-16, 2019. *Crossref,* https://doi.org/10.46253/j.mr.v2i1.a2
[31] Zeng Runhua, and Zhang Shuqun, "Improving Speech Emotion Recognition Method of Convolutional Neural Network" *International Journal of Recent Engineering Science*, vol. 5, no. 3, pp. 1-7, 2018. *Crossref,* https://doi.org/10.14445/23497157/IJRES-V5I3P101
[32] R. Tamilaruvi et al., "Brain Tumor Detection in MRI Images using Convolutional Neural Network Technique," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 9, no. 12, pp. 198-208, 2022. *Crossref,* https://doi.org/10.14445/23488379/IJEEE-V9I12P118
[33] T. Madhubala, R. Umagandhi, and P. Sathiamurthi, "Diabetes Prediction using Improved Artificial Neural Network using Multilayer Perceptron," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 9, no. 12, pp. 167-179, 2022. *Crossref,* https://doi.org/10.14445/23488379/IJEEE-V9I12P115
[34] S. Vijitha, and S. N. Sreelaja, "Modified Leading Diagonal Sorting with Probabilistic Visual Cryptography for Secure Medical Image Transmission," *Journal of Computational Science and Intelligent Technologies*, vol. 3, no. 3, pp. 1-13, 2022. *Crossref,* https://doi.org/10.53409/MNAA/JCSIT/e202203030113
[35] G. N. Srikanth, and M. K. Venkatesha, "Performance Analysis of AI Models for Audio Digit Utterance Detection," *Journal of Computational Science and Intelligent Technologies*, vol. 3, no. 3, pp. 14-29, 2022. *Crossref,* https://doi.org/10.53409/MNAA/JCSIT/e202203031429
[36] Wentao Yu, Steffen Zeiler, and Dorothea Kolossa, "Multimodal Integration for Large-Vocabulary Audio-Visual Speech Recognition," *Proceedings 2020 28th European Signal Processing Conference*, pp. 341-345, 2021. *Crossref,* https://doi.org/10.23919/Eusipco47968.2020.9287841
[37] M. Lievin, and F. Luthon, "Lip Features Automatic Extraction," *Proceedings 1998 International Conference on Image Processing, ICIP98 (Cat. No. 98CB36269)*, pp. 168-172, 1998. *Crossref,* https://doi.org/10.1109/ICIP.1998.727160
[38] Naser Karimi, and Khosro Khandani, "Social Optimization Algorithm with Application to Economic Dispatch Problem," *International Transactions on Electrical Energy Systems*, vol. 30, no. 11, p. e12593, 2020. *Crossref,* https://doi.org/10.1002/2050-7038.12593
[39] Zhicong Chen et al., "Deep Residual Network Based Fault Detection and Diagnosis of Photovoltaic Arrays Using Current-Voltage Curves and Ambient Conditions," *Energy Conversion and Management*, vol. 198, p. 111793, 2019. *Crossref,* https://doi.org/10.1016/j.enconman.2019.111793