Research Article

International Journal of Advanced Technology and Engineering Exploration, Vol 11(119)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2024.111100169

Analyzing lip dynamics using sparrow search optimized BiLSTM classifier

Shilpa Sonawane[1, 2*] and P. Malathi[3]
Research Scholar, Department of Electronics and Telecommunication, D.Y. Patil College of Engineering, Research Center, Akurdi, Pune, India[1]
Assistant Professor, Department of Electronics and Telecommunication, JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune, India[2]
Professor, Department of Electronics and Telecommunication, D. Y. Patil College of Engineering, Akurdi, Pune, India[3]

Received: 15-February-2024; Revised: 22-October-2024; Accepted: 25-October-2024

Abstract

Applications involving voice-based automatic speech recognition (ASR) have recently gained popularity. The voice-based applications fail in noisy backgrounds, overlapping speeches, and when the speech signal is completely distorted. Speech information can be recovered from the mouth region and facial emotions. The effective solution over ASR is visual speech synthesis (VSS) as it provides information about the utterance of the word from lip dynamics. The proposed methodology aims to generate speech directly from lip motion without text as an intermediate representation. A visual-voice embedding is introduced to store vital acoustic knowledge, enabling the production of audio from different speakers. The proposed sparrow search optimized bidirectional long short-term memory (BiLSTM) model takes input from lip movements and relative acoustic information, which are utilized during training. Our major contributions are: (1) suggested the use of visual voice embedding that provides additional audio information and enhances the visual aspects, thus generating superior speech from lip movements (2) the sparrow search algorithm (SSA) is employed to optimize the search for the best solution in generating audio samples from the search space, aiming to reduce loss (3) an autoregression model is proposed to produce speech from silent video without need of transcription of audio. The effectiveness of the model is checked on the GRID corpus. The performance analysis of the model is conducted by comparison between generated speech and ground truth signals concerning mean squared error (MSE), root mean square error (RMSE), signal to noise ratio (SNR), short time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ). It is observed that the proposed methodology outperforms in terms of PESQ and STOI parameters. The PESQ score shows a significant improvement of 4.06 over the generative adversarial network (GAN), while the STOI score improves by 0.202.

Keywords

Visual speech synthesis, Automatic speech recognition, Lip dynamics, Sparrow search algorithm, Bidirectional long short-term memory.

1.Introduction

In computer vision, the process of generating speech by identifying the lip motion of a speaking person without any audio input is known as visual speech synthesis (VSS). In recent years, many human-computer interaction technologies have been built using VSS technology.

VSS has huge practical potential in biometric identification [1], surveillance [1], silent dictation to smartphones in public spaces [1–3], silent passwords [4–6], Silent-movie processing [1, 7], application for speech-impaired people [1, 8–11], development of apps to assist hearing-impaired people based on lip appearances [1, 12]. It plays a significant role in noisy places where the audio signal is difficult to understand or audio signal is corrupted. If audio in videos is of poor quality, speech-to-text identification is difficult [13]. In these situations, VSS systems are helpful to obtain correct results. It helps people who suffer from speech disorders. For such people, this application provides a voice. One of the key

*Author for correspondence

functions of VSS is to set passwords and assignments of locks for numerous applications.

The domain of lip-to-speech (L2S) synthesis faces many difficulties.VSS is a difficult task for machines as some alphabets like b, d, p, m, and t have similar patterns of lip shape and motion while pronouncing them. The similarity in speech patterns causes uncertainty and confusion [14]. It is observed that each person speaks at a different rate. Hence it is difficult to notice so many speaking movements. It is difficult to capture visual information as normal speech is spoken too quickly. The early research concentrated on digits and alphabets datasets [15–17] which was an insufficient solution for continuous speech identification. The digit and alphabet datasets have a limited number of classes but it is helpful to evaluate and test the performance of the model as its time complexity is decreased as the number of trainable parameters are less for such datasets. The limited datasets lead to many challenges in handling real-world problems in target identification such as words, phrases, continuous words, and single sounds [18]. There are difficulties in handling different poses of speakers, languages, and backgrounds. In real situations, there are chances that the images appear from multiple angles and not from the front side only. The researchers worked on multi-view datasets which have many parameters like number of classes, number of speakers, multi-view of images for digit [19], phrase [19], word [20], and sentence [21] inputs.

To address these challenges, prior work in the VSS domain has concentrated on improvement in models with different features of lip and audio signals and the learning relationship between them. The earlier work has employed a convolutional neural network (CNN) based encoder-decoder architecture which works on two inputs, face images and the corresponding optical flow of the frame. The fully connected (FC) layers in the decoder produce a mel-scale spectrogram. Another work has been proposed which is based on an autoencoder with Gaussian noise to generate natural sound [22]. A GAN-based methodology was developed that generates audio directly from video for seen and unseen speakers. This research has been performed on GRID and TIMIT datasets. The previous study presented promising results on limited vocabulary, small dataset, and single speaker; they could not work correctly on the multi-speaker dataset as the model has not considered the multiple speaker lip samples. In such a scenario, the model needs to train again on multiple speaker datasets and this is time-consuming process [22]. A Lip2Wav methodology worked on unconstrained and large vocabulary datasets to handle real-world issues.

The achievement of earlier VSS models for speech generation is non-realistic since studies concentrated on the association between lip-to-text mapping, even though lip-to-text is not a one-to-one mapping. Audio-visual fusion is a difficult task. The proposed methodology aims to improve the signal to noise ratio (SNR) of generated speech to maintain the quality and intelligibility of generated speech with the assistance of lip motion, thus is most suitable when speech signal is absent. The motivation of the suggested system is to help both hearing impairment and speech-impaired people. So, hearing impairment receives text messages and speech impairment receives speech signals, enabling proper communication. The primary goal of the work is to synthesize sentence-level sequences with a deep neural network (DNN) using mouth region characteristics. The aim is to design a model with a smaller number of layers to reduce computation. The generated speech signal is then converted into text.

The primary contributions of this work can be summed up as follows: (1) a joint visual-audio collaborative model is developed to learn cross-modal features;(2) sparrow search optimization (SSA) is employed to achieve optimum solution to find audio samples in such a way to minimize error in original and generated speech signal; (3) a lip-to-audio sample generation model is designed without intermediate stage of text generation to reduce computations; (4) an autoregression model is proposed to produce speech from mute video frames without any human annotations.

The research work is compiled in six sections. Section 2 highlights previous research work on L2S reconstruction. Our approach and model architecture are discussed in section 3. In section 4, the results of the suggested approach are presented. The discussion of the model is described in section 5. The section 6 includes the conclusion of the research work.

2.Literature review

In recent years, many advanced technologies are based on deep learning models. The earlier applications based on voice recognition are replaced by vision-based characteristics since voice-based applications need silent zones. The vision-based applications were initially developed with a handful of visual characteristics [23]. Automatic sign language based on parametric representation was implemented. They presented geometry-based and pixel-based representations of visual speech. A total of 56 features were used for training in Russian sign language recognition. Russian sign language has many overlapping movements of hands covering the lip region. In such overlapping cases, the lip detection algorithm leads to error. It is hard to achieve an optimal recognition rate. They suggested statistical approaches like the hidden markov model (HMM) or DNN or automatic sign language recognition models [23]. A model was developed by focusing on both audio and visual characteristics [24]. In an implementation, mel frequency cepstral coefficients (MFCCs) of the audio signal are used along with visual cues extracted using CNN. Both features are combined and forwarded to the HMM model to recognize isolated words. The experimentation results attain reliable audio visual speech recognition (AVSR) performance. They suggested testing the AVSR architecture in a real-world environment to handle dynamic changes.

The performance is enhanced by later technologies like HMM [25] and a hybrid combination of DNN-HMM. A hybrid DNN-HMM model was developed for lip reading. In DNN, backpropagation is used to reduce cross-entropy between the target HMM state and predicted output. Eigen lip feature of the mouth region of interest (ROI) is found by principal component analysis (PCA). The accuracy is computed at the word level for the TCD-TIMIT dataset. The accuracy is enhanced from 4% to around 51% in speaker-independent scenarios. It is found that DNNs are good configurations for processing of complex density of lip features while they cannot simulate that variation by identification [25]. In the same DNN-HMM configuration, discrete cosine transform (DCT) and Eigen lip features are combined to represent mouth ROI images. The developed lip-reading configuration employs either 13 viseme or 38 phoneme units to simulate visual speech. The phoneme-based lip-reading model shows better performance than that of the viseme-based approach. The speech reconstruction frameworks based on the gaussian mixture model (GMM) and DNN were developed by employing the active appearance model (AAM), 2D-DCT features, and 8th-order linear predictive coefficients (LPC) features. The word accuracy was computed for two approaches for reconstructed speech. The accuracy obtained using visual features alone is 49.02%. LPC features were employed for the estimation of time-frequency information from visual cues. They suggested focusing on improving audio information and testing the model's performance in a less restricted environment [26]. A lip-reading system was presented based on classification methodologies like HMM, support vector machine (SVM), and k-nearest neighbour (KNN) [27]. The appearance and motion characteristics of the mouth region are found in mouth ROI. All characteristics are bundled into compact feature vectors and fed to classifiers. The evaluation of the system is shown expressed by accuracy using MIRACL-VC1, CUAVE, OuluVS, and BIWI Kinect head pose datasets. In this research work, speakers are not permitted to move their heads in the direction of yaw and pitch. The upcoming work is to focus on speech segment identification in continuous video streams [27].

In recent years, the effectiveness has improved further using DNNs like CNN, long short-term memory (LSTM), gated recurrent unit (GRU), and encoder-decoder architecture. A LipNet model was presented which is the first end-to-end model that finds the entire sentence from the sequence of images of mouth ROI. CNN with GRU was employed to find spatio-temporal information. The analysis of LipNet was performed for Word Error Rate (WER) and character error rate (CER) on the GRID corpus. The average WER obtained is 47.7%. As the employed dataset is limited, they need to evaluate the performance on a larger dataset. The Lipper model was configured to reconstruct audio from visual information which is at multiple angles. It is constructed by a combination of spatiotemporal CNN and bidirectional GRU. The experiment was conducted on the OuluVS2 database to evaluate the performance in terms of perceptual evaluation of speech quality (PESQ). It is observed that the average PESQ obtained is 2.086 for all views. The limitation of the system is that generated audio seems artificial as the model focused on mouth ROI only. Therefore, the model needs improvement to make the generated speech more realistic [28].

A Lip2Wav model was developed based on spatial and temporal face coder and voice decoder. A face encoder is a stack of three-dimensional (3-D) convolution employed to find spatiotemporal attributes of lip movements. The attention-based decoder is developed to reconstruct speech from visual attributes. The reconstruction loss is reduced between the generated MFCC and the MFCC of the original signal. The model is tested on GRID and TCD-TIMIT datasets in terms of PESQ and short time objective intelligibility (STOI) metrics. STOI and PESQ obtained are 0.731 and 1.772 respectively [29]. An auto-encoder and decoder were implemented to reconstruct speech signals from lip motion. A deep network is employed to extract speech-related features from the mouth region. The features are encoded and decoded using a pre-trained auto-encoder. The quality of the coded-decoded signal is computed in terms of PESQ. The experimentation was conducted on GRID corpus and obtained PESQ using the auto-encoder network is 2.76. Since the work was conducted on a limited dataset, they suggested collecting larger datasets that include emotions and developing an end-to-end trainable model to directly reconstruct audio [30]. Later on, a system is implemented for lip-reading based on the GRU network, hybrid network like CNN followed by LSTM. The features are computed using pre-trained networks like residual neural network-50(ResNet-50), visual geometry group19 (VGG) and CNN. LSTM is used for learning the time sequence information. The experiment is conducted on a short video uttering a single word as a digit. In experimentation, only one word at a time is detected. As digits were focused on experimentation, a greater number of words is expected to be identified from an input video sequence at a time.

A model was designed to help hearing-impaired people and recognize speech in noisy places [31]. Audio and visual information are integrated and passed to deep learning architectures to recognize speech from lip motion. The accuracy obtained is 95% while the error rate achieved is 6.59%. They suggested to improve WER by making use of bidirectional encoder representations from transformers (BERT) based language model. A visual speech recognition (VSR) model was developed to recognize words based on a spatiotemporal deep learning network [32]. The experiment was conducted on a BBC TV broadcast dataset of size 500 target words. A residual network is employed to identify visual cues and passed to bidirectional LSTM for classification. The word level accuracy obtained is 83% and it is a 6.8% improvement over the attentional encoder-decoder network. A VSR framework was presented which makes use of audio information. The visual audio memory (VAM) concept was implemented which learns audio information of video in a memory network using corresponding visual information. VAM includes two components namely lip-visual key and voice attributes. The mouth-visual pair is used to remember the position of voice attributes. Therefore, VAM provides more information on lip motion and its corresponding audio features and thus improves the performance of VSR [33].

A novel flash attention GAN (FA-GAN) is developed to boost speech recognition accuracy. Swin transformer is implemented to find both local and global visual cues from video sequences. The computational complexity and visual ambiguity are minimized by employing an attention strategy. Experimentation is conducted on GRID and Chinese continuous visual speech (CN-CVS) datasets. FA-GAN model achieves 0.614 STOI, 0.580 extended short-time objective intelligibility (ESTOI), and 1.772 PESQ for the Chinese dataset. The FA-GAN framework still faces difficulties, such as greater variation in lip movements and the intrinsic complexity of Chinese pronunciation and speech [34]. The LipSound2 model is introduced which utilizes an encoder-decoder framework with an attention technique to produce a mel spectrogram from video frames without human annotations. The encoder consists of multitask CNN to identify face landmarks followed by bidirectional long short-term memory (BiLSTM) to extract temporal dependencies. The location-aware attention module is used to map video features to the spectrogram. The model is tested on GRID and TCD-TIMIT datasets. The performance parameters such as ESTOI and PESQ for the GRID dataset are 0.592 and 2.328 respectively. The suggested LipSound2 model is trained on the large-scale dataset, it performs both lip reading and speech prediction tasks properly, but it still throws errors in generated speech due to the visual similarity in pronunciation [35].

One of the challenges is to develop a lip-reading model for low-resource languages because it does not have video-text paired data to train the model. To address this problem suggested methodology learn general speech knowledge and language-specific knowledge from audio-text paired data. By integrating both the learned general speech knowledge and language-specific knowledge, the lip-reading model is developed for low-resource languages. Lip reading performance is evaluated on m TEDx-PT in terms of WER. It is about 69.33% for the English language. The specific audio features are saved in the language memory (LM) decoder hence it provides more information when it is combined with general speech knowledge. The performances of automatic speech recognition (ASR) in LM decoder are not superior to those of the ASR pre-train model [36]. A communication system using deep-learning is introduced for speech transmission and synthesis. The transmitter consists of the feature and channel encoder. The features required for speech recognition are extracted by the channel encoder at the transmitter. The receiver has a channel and feature decoder to reconstruct text-related features and recognize the final text which is converted into audio by speech synthesis module. Tacotron is used to reconstruct speech samples from spectrograms. The deep semantic communication for the speech synthesis model demonstrates superior speech recovery than the speech, feature, and text transceiver. The results show that the reconstructed speech signals using a speech transceiver are not acceptable when SNR is below 4dB under the additive white gaussian noise (AWGN) channel [37]. A novel approach for automatic lip-reading is implemented by focusing on the intensity level of the speaker's emotion. The suggested approach uses visual speech data to detect a person's emotion type and intensity level. 3-D CNN and BiLSTM are utilized to tackle emotional speech lip-reading. The suggested method improves the results by up to 8.2% in terms of accuracy due to consideration of the intensity of pronounced audio-visual speech [38]. A visual text-to-speech(TTS) network is designed to generate sound from the input silent video stream. A pre-trained lip-to-text network is used to extract visual features and text. The phonemes are retrieved for each video frame using the attention mechanism. These are then up-sampled and decoded into mel spectrograms. Finally, mel spectrograms are converted into speech signals using a vocoder. Using the suggested approach, the PESQ obtained is 1.47 and the STOI is 0.655. The limitation of the developed lip-to-text model is that the model is trained using text supervision [39].

An authentication system is developed to validate passwords. The system makes use of facial recognition and an individual's temporal facial features while speaking passwords in any language. This application was tested on the MIRACL VC1 dataset and they found difficulty in the development of authentication systems for personal computers and mobile devices concerning the amount of time spent in testing the system. A novel LipVoicer method is introduced to generate high-quality speech from silent video. In this, the lip-reading model is developed which produces text from the video. In the next phase, ASR uses predicted text to generate a mel spectrogram. The raw audio is generated using the DiffWave neural vocoder. The model is evaluated on the challenging LRS2 and LRS3 datasets. Although the suggested lipreading model is trustworthy, false text may be inserted by "bad-faith actors" [40]. A model named video-to-speech synthesis is designed that uses video and audio inputs during both training and inference time. In the first step, a pre-trained video-to-speech model is developed to predict missing audio. In the second step, predicted audio along with video feed to audio-visual-to-audio model to produce speech signal. The system is tested on GRID and TIMIT datasets. It is observed that the PESQ obtained is 1.45 and STOI is 0.598 for the GRID dataset [41].

Due to several obstacles, most VSS models have not been able to fulfil real-world needs despite years of development. The objective of VSS is more subjective and multifaceted than VSR. A video of a speaking person consists of coupled information like identity and motion-related information. The "uncanny valley effect" causes people to see a synthetic face that looks almost human but not quite perfect. Extensive details regarding the source identity are typically found in the driving source. Thus, to avoid corruption during the target identity synthesis process, the important task is how to remove the identity information from the driving source. The most of current techniques are limited to certain target identities. The absence of identity generalization is thus a major difficulty. The motion-related challenges have to consider intrinsic and extrinsic motions like head movements, facial emotions, lip movements, background motion, etc. Integrating synthesized speech with video frames is also an issue as accurate temporal matching is needed. In addition to the above obstacles, there are data-related issues in VSS. The existing dataset has scant annotations and is of limited size. Visual speech is a natural signal and has multiple dimensions like speaking style, speaking rate, speaker age, language, gender, emotions, etc. There are still challenges for the development of deep learning models to consider multiple aspects of audio-visual datasets. It is found from the literature that generated speech lacks emotions and looks robotic as in visual information, excitation source information is absent. The previous research study shows that the models implemented are complex in nature and more training parameters are required to generate intelligible speech signals. The motivation for exploration is to construct a simpler model with fewer computations. The research aims to synthesize intelligible speech from silent video from multiple speakers. In the proposed technique, the implemented model is 11 layers deep, and generally, the Adam optimizer is used in previous research work. In this approach, SSA is used to identify the best voice sample from the search space.

3.Method

The objective of the proposed methodology is to produce a speech signal from the lip dynamics of the speakers as the input is a silent video signal. The methodology uses BiLSTM classifier to map the visual information into voice samples. The suggested system uses SSA-based BiLSTM and ResNet-50 architecture. The steps of VSS work are presented in *Figure 1*.

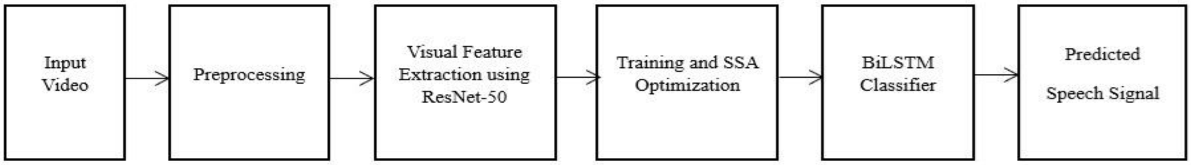

Input Video
Preprocessing
Visual Feature Extraction using ResNet-50
Training and SSA Optimization
BiLSTM Classifier
Predicted Speech Signal

Figure 1 Methodology of VSS

The input for the suggested technique is a silent video signal. The input video has a sequence of video frames. Each video frame is pre-processed to detect the face region as the aim is to localize the mouth region. The objective of the proposed methodology is to generate speech signals from the mouth region, so there is a need to find visual characteristics. The characteristics of the mouth region are computed by using the ResNet-50 model. The training process uses visual features and corresponding audio frames to learn the relation between them. The SSA is implemented to find the best solution to recover an audio sample from the search space. In the end, a BiLSTM classifier is used to produce a speech signal.

3.1Dataset

The proposed system is analyzed on the GRID audio-visual dataset [41]. It consists of short sentences recorded by 34 speakers. The male speakers are 16 and the female speakers are 18. A total of 1000 sentences are recorded by each speaker. All the videos are available in full frontal pose. Each video has a duration of three seconds with a frame rate of 25 frames per second. The video files are available in two types: high definition (720 ×576) and normal quality (360 × 288). The duration of the audio and video track is the same with a 50 kHz sample frequency. It makes use of 6 words from a pre-defined dictionary to frame English sentences as shown in *Table 1*.

Table 1 GRID dataset word categories

Sr. No.	Categories	Candidate words
1	Command	bin, lay, place, set
2	Color	blue, green, red, white
3	Preposition	at, by, in, with
4	Letter	A, . . . Z (W excluded) Digit zero, . . ., nine

3.2Input video pre-processing

The input video requires to be pre-processed to separate video frames. The procedure to get the set of frames that is necessary for feature extraction is presented in *Figure 2*.

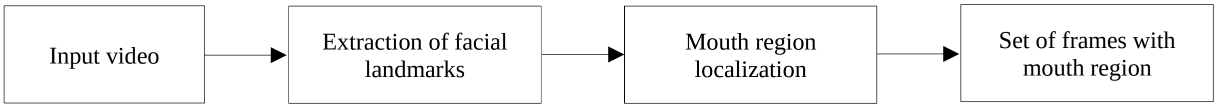

Input video
Extraction of facial landmarks
Mouth region localization
Set of frames with mouth region

Figure 2 Input video pre-processing

The input video is of the speaker uttering a sentence. There is a need to pre-process individual video frames to find useful information from them. Therefore, extracting the frames from the video is the initial step. Each video has 25 frames per second and it is of 3-second duration. It is observed from the video, that there is redundant information about the movements of the speaker. To reduce processing time and memory to store movements, need to ignore some redundant information. It is also helpful to improve training speed. The proposed methodology does not utilize every frame from the video sequence to remove redundant information rather it extracts the key frames from the video sequence at intervals of three. The complete video is now transformed into a sequence of 25 frames after one frame is chosen at an interval of three. The selected frame is further processed to detect facial landmarks. An essential step in data preprocessing is to maintain uniformity of numeric data. Due to uniformity, features with higher values than the other features don't dominate. When all features are scaled to the same value, it gets simpler to recognize connections with different features. Thus, normalized data help to improve both the accuracy and performance of a model. The input signal is normalized by dividing each sample of the signal by its greatest value**.**

3.3Extraction of the mouth region

The well-known face extraction algorithm by Dlib is employed to extract 68 facial landmarks [42]. The histogram of oriented gradients (HOG) of the frame is computed. The corners and edges are identified using HOG features since Gradients are usually larger around edges and corners and thus helpful to identify the regions. The facial landmarks are shown in *Figure 3*. The points 49 to 68 are used to extract the mouth region. The points 49, 55, 52, and 58 were taken into consideration when cropping the mouth area. The mouth ROI is of size 51×126×3.

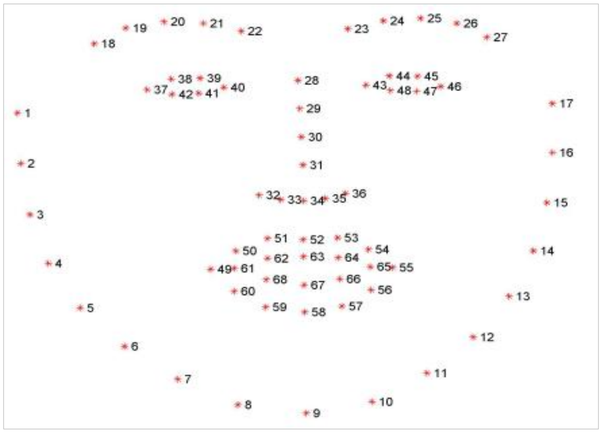


Figure 3 Facial landmarks using Dlib

3.4Visual feature extraction

To generate speech signals from lip movements, it is necessary to find the correlation between lip motion and voice sample. The visual characteristics are useful for finding voice samples. The visual cues are extracted using ResNet-50 CNN architecture. It provides the collection of spatial information for every input frame. A specific kind of CNN known by the ResNet was first presented in 2016 [43]. The ResNet architecture is based on two basic design rules. The first is, based on the measurement of the output characteristic vector, each layer has a similar filter size. Second, to keep up the temporal complexity of each layer, if the dimension of the feature vector is halved then the filter count is doubled. ResNet-50 model uses a bottleneck design. A bottleneck residual block uses 1×1 convolution which lowers the network attributes and matrix operations. This authorizes each layer to be trained considerably faster. Instead of using two layers, it uses a stack of three layers as shown in *Figure 4*. It has a total of 50 layers altogether in which 48 layers perform convolution operation, one MaxPool, and one average pool layer. The proposed methodology takes into account the network up to ResNet-50's FC layer. The features extracted from the FC1000 layer are used to find voice samples. ResNet-50 network takes an image input size of 224-by-224. It is a pretrained neural network trained on over 14 million photos from the ImageNet database. It is capable of classifying images into 1000 different object categories. Thus, it extracts rich feature representations for a variety of images. *Table 2* provides detailed information about the layers used in it. A first layer with 7×7 kernel convolution alongside 64 other kernels with a 2-sized stride. The next layer is the max pooling layer with a 2-sized stride. The next convolution includes three kernels: 1 × 1,64, 3 × 3, 64, and 1 × 1,256. These three layers are repeated three times in total, thus nine layers in this stage. The kernels of 1×1,128 are added next, followed by 3 × 3,128 and, finally, 1 × 1,512 resulting in 3 layers which are repeated 4 times producing a total of 12 layers in this phase. Then again three kernels are 1 × 1, 256, 3 × 3, 256, and 1 × 1, 1024, and this is repeated 6 times giving a total of 18 layers. After that, there is a kernel of 1 × 1, 512, 3 × 3,512, and 1 × 1,2048 and this was repeated 3 times giving us a total of 9 layers. Then applied average pool operation was followed by a FC layer with 1000 nodes and a SoftMax function at the end, providing output.

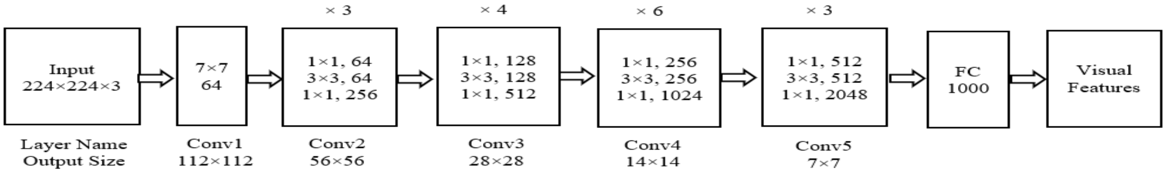


Figure 4 ResNet-50 architecture

Table 2 Configuration of ResNet-50

Layer name	Output size	Resnet-50	Multiplier	No. of layers
		Input=224 × 224 × 3		
Conv1	112 × 112	7 × 7, 64, stride 2		1
Conv2	56 × 56	3 × 3 max pool, stride 2		9
		1 × 1, 64	3	
		3 × 3, 64	3	
		1 × 1, 256	3	
Conv3	28 × 28	1 × 1, 128	4	12
		3 × 3, 128	4	
		1 × 1, 512	4	
Conv4	14 × 14	1 × 1, 256	6	18
		3 × 3, 256	6	
		1 × 1, 1024	6	
conv5	7 × 7	1 × 1, 512	3	9
		3 × 3, 512	3	
		1 × 1, 2048	3	
	1 x 1	avg pool, 1000-dimensionn FC		
		Total Layers		50

3.5Sparrow search optimized BiLSTM classifier

The set of feature vectors extracted from ResNet-50 is processed by SSA-optimized BiLSTM. The features are flattened and passed through three stages of BiLSTM followed by a rectified linear unit (ReLU) as shown in *Figure 5*. Then the information is passed to a FC layer. Sparrow search optimization is used to find appropriate audio samples from available search pool [44]. The output is forwarded to the regression layer to find the correct voice samples.

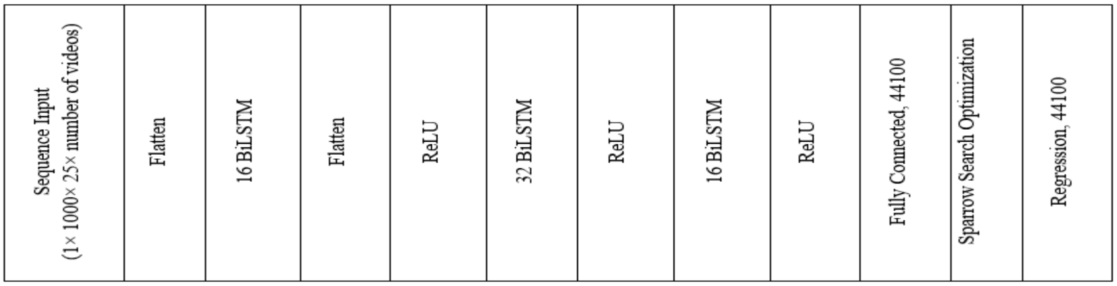


Figure 5 Sparrow search optimized BiLSTM classifier

In DNN, the flattened layer is frequently introduced to convert the multidimensional input into a one-dimensional vector. Activation functions help hidden nodes to generate more desired output by introducing nonlinearity to the model. It also removes the vanishing gradients problem from the model. The ReLU activation carries out a nonlinear threshold operation. Any input value less than zero is set to zero by using ReLU. Each neuron in a FC layer receives the information from the previous layer neuron and generates the output based on weights, bias, and activation functions. The model learns the relationship between input and output data and converts the learned features into the format so that they may be applied to the regression layer. The regression layer is employed to predict the continuous value of speech signal from previous information of speech sample. The configuration of BiLSTM is shown in *Table 3* in which 'S' denotes spatial, C is for channel, B is batch and T represents time.

Table 3 Configuration of BiLSTM classifier

S. No.	Layer type	Activations	Learnable sizes
1	Sequence Input	1000(S) × 25(C) × 1(B) × 1(T)	-
2	Flatten	2500 (C) × 1(B) × 1(T)	-
3	BiLSTM (16 hidden units)	32 (C) × 1(B) × 1(T)	Input Weights 128×25000 Recurrent Weights 128×16

S. No.	Layer type	Activations	Learnable sizes
			Bias 128×1
4	Flatten	32 (C) × 1(B) × 1(T)	-
5	ReLU	32 (C) × 1(B) × 1(T)	-
6	BiLSTM (32 hidden units)	64 (C) × 1(B) × 1(T)	Input Weights 256×32 Recurrent Weights 256×32 Bias 256×1
7	ReLU	64 (C) × 1(B) × 1(T)	
8	BiLSTM (16 hidden units)	32 (C) × 1(B) × 1(T)	Input Weights 128×64 Recurrent Weights 128×16 Bias 128×1
9	ReLU	32 (C) × 1(B) × 1(T)	
10	FC (44100)	44100 (C) × 1(B) × 1(T)	Weights 44100×32 Bias 44100×1
11	Regression Output	44100 (C) × 1(B) × 1(T)	-

3.5.1Sparrow search optimization algorithm

The mathematical model of SSA [44] is described in this section. In SSA, the matrix can be used to represent the positions of the sparrows as shown in Equation 1:

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & \cdots & P_{1,D} \\ P_{2,1} & P_{2,2} & \cdots & \cdots & P_{2,D} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{N,1} & P_{N,2} & \cdots & \cdots & P_{N,D} \end{bmatrix} \qquad (1)$$

In Equation 1, N shows the total count of sparrows, and D is the dimensions of the feature vector. The overall sparrows' fitness value is represented in Equation 2.

$$F_P = \begin{bmatrix} f([P_{1,1} & P_{1,2} & \cdots & \cdots & P_{1,D})] \\ f([P_{2,1} & P_{2,2} & \cdots & \cdots & P_{2,D})] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f([P_{N,1} & P_{N,2} & \cdots & \cdots & P_{N,D})] \end{bmatrix} \qquad (2)$$

Where F_P represents the fitness function and the number in every row indicates the fitness value of each sparrow. In the search process for food, producers having greater fitness value are given priority. Producers have to look for food and direct the flow of sparrows. The producers' spot is upgraded in each iteration by using Equation 3.

$$P_j^{itr+1} = \begin{cases} P_{i,j}^{itr}.\exp\left(\frac{-i}{\propto.iter_{max}}\right) & if\ R_2 < ST \\ P_{i,j}^{itr} + Q.L & if\ R_2 \geq ST \end{cases} \qquad (3)$$

Where j=1, 2… D and itr is the current iteration, P_j^{itr+1} indicates the element of the jth column and ith sparrow at a specific iteration. $iter_{max}$ is maximum iterations, α is the arbitrary numeral between 0 and 1, and R_2 is the alarm value ($R_2 \in [0, 1]$) and ST is a safety threshold between 0.5 and 1.0. Q indicates arbitrary numeral follows a normal distribution. M represents a matrix of element 1 of size 1 × D. The producer moves to the wide search mode when the first condition, R2 < ST, is met, indicating that there are no predators in the area. The other condition R2 ≥ ST shows that some sparrows have detected the predator, in which case each of the sparrows must go to safer locations. Once the scroungers notice that the producer has discovered nice food, they change their present location to get the food. The position of scroungers is updated using Equation 4.

$$P_j^{itr+1} = \begin{cases} Q.\exp\left(\frac{P_{worst}^{itr} - P_{i,j}^{itr}}{\propto.iter_{max}}\right) & if\ \mathrm{i} > \mathrm{N}/2 \\ P_p^{itr+1} + \left|P_{i,j}^{itr} - P_p^{itr+1}\right|.Z^{+}.\mathrm{M} & \text{otherwise} \end{cases} \qquad (4)$$

where P_P represents the producer's ideal place. The current global worst location is shown by P_{worst}. The matrix Z is of size 1 × D which includes -1 or 1, and $Z+ = Z^T (ZZ^T)$-1. The ith scrounger has the least fitness value which indicates it is probably to be starving when i becomes greater than N/2.

When sparrows sense danger, they instantly go toward the safe region to take up a secure place, and sparrows at the centre move randomly to come closer to each other. The mathematical model to represent this is described in Equation 5.

$$P_j^{itr+1} = \begin{cases} P_{best}^{itr} + \beta.\left|P_{i,j}^{itr} - P_{best}^{itr}\right| & if\ f_i > f_g \\ P_{i,j}^{itr} + \mathrm{K}.\left(\frac{\left|P_{i,j}^{itr} - P_{worst}^{itr}\right|}{(f_i - f_w) + \varepsilon}\right) & if\ f_i == f_g \end{cases} \qquad (5)$$

Where P_{best} shows the current global ideal place. β represents step size adjustment factor which follows normal distribution with zero average and variance of one. Here f_i is the fitness value of the present sparrow. f_g and f_w are the current global top and least fitness values, respectively. ε is the lowest integer to avoid a divide-by-zero error. The condition f_i becomes greater than f_g indicating that the sparrow is at the boundary of the group. P_{best} denotes the safe centre location of the sparrows. The condition $f_i == f_g$ represents that sparrows need to get closer to each other because they are aware of the risk. The route of sparrows' motion is indicated by K and it is in the range of -1 to 1. The optimization parameters used in SSA are mentioned in *Table 4.*

Algorithm: Flow of SSA

```
N = population of sparrow.
NP = population of producers who recognize the danger.
NS = population sparrows who recognize the danger.
itr=1
 while (count < Max number of iterations)
          Sort the fitness metrics in order of preference and discover the current top and bottom individuals.
          $R_2$ = random (1)
          for i = 1: NP
               Upgrade the position of the sparrow using equation (3).
          end for
         for i = (NP + 1): N
              Upgrade the position of the sparrow using equation (4).
         end for
         for k = 1: NS
              Upgrade the position of the sparrow using equation (5).
         end for
         if the new position is better than the previous one, change it.
            itr = itr + 1
end while
return $P_{best}$, $f_g$.
```

Table 4 SSA optimization parameters

S. No.	SSA optimization parameter	Parameter value
1	Number of sparrows	100
2	Number of producers	20
3	Maximum iterations	1
4	Alpha	0 to 1 (randomly selected)
5	Beta	0 to 1 (randomly selected)
6	Safety threshold	0.8

4.Results

The experimentation is conducted on the GRID audiovisual dataset. Our network implementation is based on the MATLAB 2024a version with an image processing, audio, and deep learning toolbox. The training of networks is an essential step to learning the relationship and dependencies between different features. Training options for the proposed model are listed in *Table 5*. Network weights and bias values are initialized using the network initialization function. The initial learning rate is set to 0.0010 with Adam optimizer. A mini-batch of 500 training samples was used. The dataset was divided into training, testing, and validation sets with proportions of 70%, 20%, and 10%, respectively. The network training stopped when the validation loss stopped decreasing around 500 epochs. The network is trained with backpropagation using root mean squared error. The BiLSTM model is introduced to learn visual features extracted from Resnet-50. The visual and audio data both are high dimensional data. The experiment is conducted on AMD Ryzen-7 5800H with Radeon Graphics with 3.20GHz along with NVIDIA RTX GPU with 4GB memory.

Table 5 Training parameters of BiLSTM

S. No.	Training parameters	Parameter value
1	Initial learning rate	0.0010
2	Batch size	500
3	Max Epoch	500
4	Loss function	Root mean square error (RMSE)
5	Solver	Adaptive moment estimation (Adam)

Training a network is the process of minimizing the loss function which is equivalent to reducing the reconstruction loss. Thus, the correlation between the sequence of lip motion and its corresponding speech in the time domain is highest. The training results for BiLSTM and BiLSTM with SSA are shown in *Tables 6* and *7* respectively. It is observed that as epoch increases RMSE and loss decrease in both classifiers.

Table 6 Training of BiLSTM classifier

Epoch	**Iteration**	**Time Elapsed (hh:mm:ss)**	**Mini-batch RMSE**	**Validation RMSE**	**Mini-batch loss**	**Validation loss**	**Base learning rate**
1	1	00:00:14	9.30	9.28	43.2129	43.0960	0.0010
50	50	00:00:39	8.94	8.94	39.9962	39.9195	0.0010
100	100	00:01:09	8.42	8.41	35.4156	35.3554	0.0010
150	150	00:01:40	7.89	7.88	31.1655	31.0544	0.0010
200	200	00:02:11	7.37	7.36	27.1927	27.1005	0.0010
250	250	00:02:43	7.15	7.13	25.5363	25.4242	0.0010
300	300	00:03:16	6.84	6.82	23.3629	23.2565	0.0010
350	350	00:03:51	6.77	6.75	22.8926	22.8000	0.0010
400	400	00:04:29	6.81	6.84	23.1817	23.3630	0.0010
450	450	00:05:06	6.78	6.76	22.9954	22.8466	0.0010
500	500	00:05:46	6.67	6.67	22.2731	22.2479	0.0010

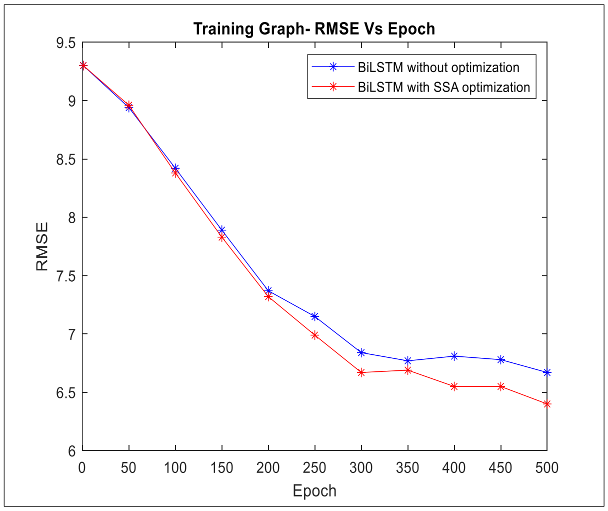
Table 7 Training of BiLSTM with SSA optimization

Epoch	**Iteration**	**Time elapsed (hh:mm:ss)**	**Mini-batch RMSE**	**Validation RMSE**	**Mini-batch loss**	**Validation loss**	**Base learning rate**
1	1	00:00:04	9.30	9.28	43.2128	43.0975	0.0010
50	50	00:00:50	8.96	8.95	40.1102	40.0365	0.0010
100	100	00:01:31	8.38	8.37	35.1473	35.0646	0.0010
150	150	00:02:11	7.83	7.82	30.6361	30.5472	0.0010
200	200	00:02:52	7.32	7.32	26.8093	26.8072	0.0010
250	250	00:03:32	6.99	6.96	24.4063	24.2019	0.0010
300	300	00:04:12	6.67	6.68	22.2305	22.3151	0.0010
350	350	00:04:53	6.69	6.68	22.3983	22.3000	0.0010
400	400	00:05:35	6.55	6.54	21.4407	21.3856	0.0010
450	450	00:06:18	6.55	6.53	21.4481	21.3374	0.0010
500	500	00:07:00	6.40	6.39	20.4623	20.4241	0.0010

The comparison between RMSE and loss for both models is shown in *Figures 6* and *7* respectively. It is found that the performance of BiLSTM with SSA is improved over BiLSTM without optimization.

The performance analysis of the model is conducted by computing parameters of the generated speech signal as per speech standard. SNR computes the proportion of undesired noise to audible speech in an audio stream. The greater values of SNR indicate better specification since they represent more useful information about the signal than the noise. RMSE is widely utilized to measure the differences between true or predicted values. A smaller RMSE is generally preferable to a greater one. PESQ specified by the international telecommunication union telecommunication standardization sector (ITU-T) P.862 standard, is a technique to test voice quality in telephone networks. PESQ is an accepted industry standard for audio quality that accounts for factors like call volume, audio clarity, background noise, clipping, and interference. The higher scores indicate better quality [45]. The short-time objective intelligibility (STOI) parameter is computed using an association between the temporal envelopes of original and generated speech. The prediction results are achieved by implementing two models namely BiLSTM without optimization and SSA optimization with BiLSTM as shown in *Table 8.*

The performance parameters obtained for SSA-BiLSTM are compared with BiLSTM without optimization. It is observed that the results are improved in SSA-BiLSTM. The generated speech signal and its spectrogram using both models are shown in *Figures 8* and *9*. The predicted speech is converted into text using instruction speech2text using Matlab.



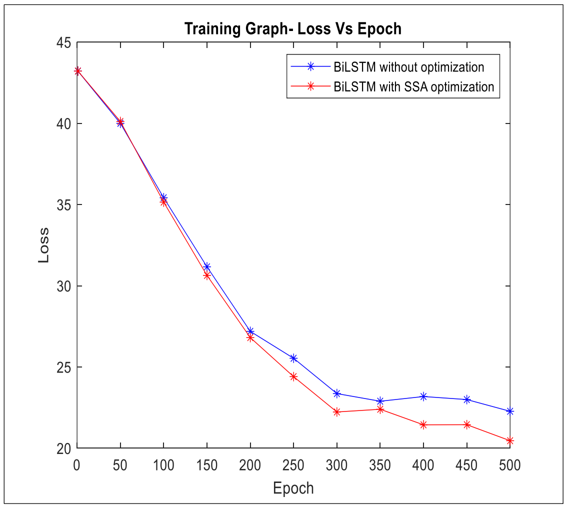


Figure 6 Comparison of training graph-RMSE vs. Epoch, **Figure 7** Comparison of training graph-loss vs. Epoch

Table 8 Performance parameters for classifiers

S. No.	Performance parameters	BiLSTM	SSA-BiLSTM
1	SNR (dB)	24.21307	24.36546
2	RMSE	23.80351	22.80351
3	PSNR (dB)	43.96546	44.28546
4	Mean squared error (MSE)	23.31148	22.31148
5	PESQ	5.421308	5.436546
6	STOI	0.728374	0.770792

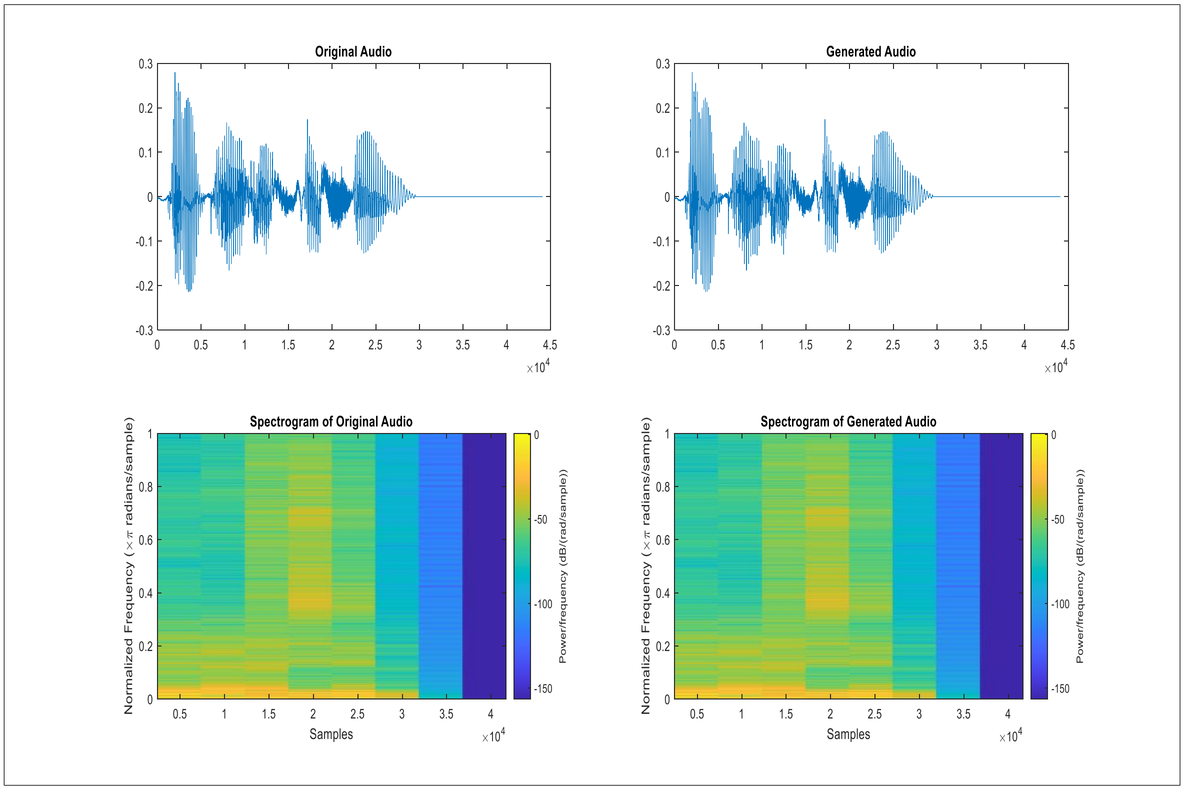


Figure 8 Predicted speech and MFCC using BiLSTM without SSA

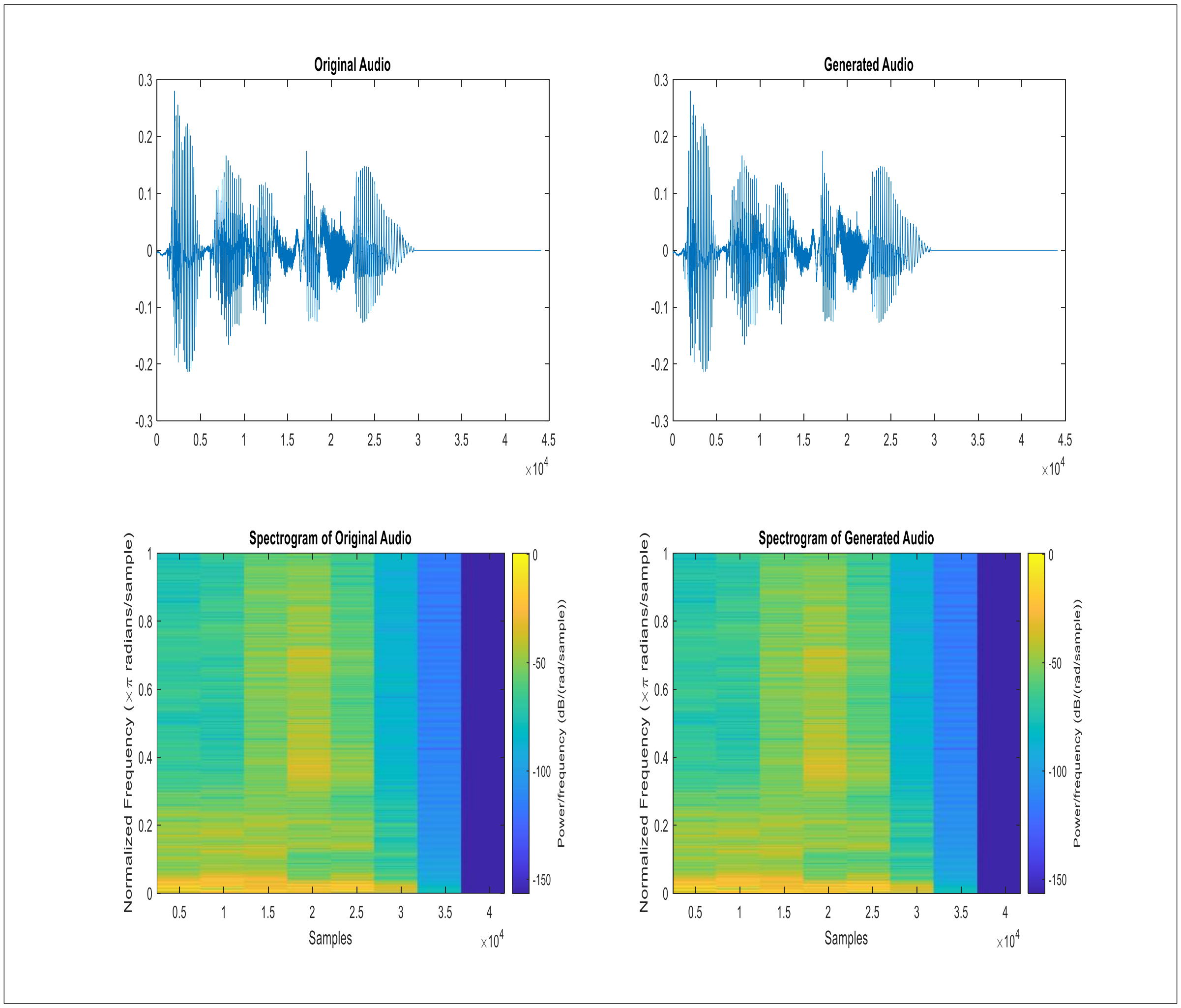


Figure 9 Predicted speech and MFCC using BiLSTM with SSA

In an experiment with the GRID dataset, as shown in *Table 9*, it is observed that the performance of the SSA-BiLSTM methodology is superior to all other comparable methods across reconstruction metrics.

Table 9 Comparison of results on the GRID dataset

S. No.	Method	PESQ	STOI
1	Lip2Wav [35]	1.772	0.731
2	End-to-end GAN [46]	1.37	0.568
3	SVTS-M [47]	1.40	0.588
4	Visual context attentional-generative adversarial network (VCA-GAN) [48]	1.43	0.589
5	Robust L2S [49]	-	0.754
6	Voice conversion-based video-to-speech (VCVTS) [50]	1.816	0.691
7	BiLSTM (proposed methodology)	5.42	0.728
8	SSA-BiLSTM (proposed methodology)	**5.43**	**0.770**

5.Discussion

The proposed system works well for speaker-dependent samples. The initial stage is the detection of the lip region. Dlib toolbox is efficient in extracting mouth ROI. It works efficiently for all videos of the GRID dataset. The crucial step is processing of entire video to extract mouth ROI and visual features. Our observations indicate that face detection and the extraction of the mouth region and visual features are time-consuming processes. The

GRID dataset includes the videos which consist of speakers with frontal poses. As observed in the literature survey, our work can be extended towards non-frontal talking face videos.

The model can generate an intelligible speech signal which is approximately similar to the original speech signal because voice embedding is done in an encoder which efficiently helps both models to learn the features and efficiently predict the speech samples. The limitation faced during testing of our model is for female speakers, it generates the speech content correctly but the voice is of the male speaker and thus leads to failure of the system because our system focused only on male speakers' data during training. This obstacle was resolved by including female samples in the training set, allowing the network to learn the characteristics of female voices. The identical dataset configuration used in Lip2Wav was considered, and the model was tested on both male and female speakers. The results demonstrate the effectiveness of the model in speaker identification and speech production as our model concentrated visual-voice pair in the training phase.

The comparison begins with earlier state-of-the-art L2S synthesis tasks, which were experimented on using the GRID dataset. *Figures 10* and *11* show the comparison of the generated speech of the proposed model with other previous methods namely Lip2Wav, VCVTS, and robust L2S for male and female speakers respectively. The *Figures 10* and *11* indicate that the generated audio waveform of the proposed model SSA-BiLSTM matches the ground truth waveform. It is also found that the Lip2Speech model produces a waveform almost identical to the real one. The speech waveform generated using Lip2Wav and VCVTS models is not quite similar to the ground truth signal. Using Lip2Wav and VCVTS, the magnitude of speech samples is less as compared to the ground truth which affects the pitch of the signal.

Table 8 shows that our model outperforms other previous methods, including state-of-the-art performance achieving 5.43, and 0.770, in PESQ and STOI scores respectively. In our approach, visual-voice embedding incorporates extensive audio information to store lip motion and match audio context when creating speech from the silent video clip. Thus, consequently producing excellent speech in both single- and multi-speaker-dependent scenarios. The proposed system works well for speaker-dependent samples. The initial stage is the detection of the lip region. Dlib toolbox is efficient in extracting mouth ROI. It works efficiently for all videos of the GRID dataset. The crucial step is processing of entire video to extract mouth ROI and visual features. Our observations indicate that face detection and the extraction of the mouth region and visual features are time-consuming processes. The GRID dataset includes the videos which consist of speakers with frontal poses. As observed in the literature survey, our work can be extended towards non-frontal talking face videos.

The performance of the proposed architecture on the English GRID dataset indicates a slight improvement in the STOI metric compared to the robust L2S methodology, with a 0.016 rise. The SSA optimized BiLSTM model outperforms comparable models in PESQ and STOI metrics.

A high-dimensional visual feature vector is constructed from multiple frame sequences of the input video signal. Thus, additional time is needed to train the network. Due to the requirement to reduce the size of the feature vector, SSA optimization is introduced for optimal feature selection. Thus, SSA optimization makes the model more effective in mapping lip movement changes to voice signals more precisely. The producers in SSA are responsible for obtaining global optimization by exploring new regions of the search space. Scroungers further enhance the search process for the best solution in the adjoining space by utilizing the currently promising locations, which improves local optimization.

The previous VSS work employed VSR with TTS, using text as an intermediate representation. Such text-based models require annotated labels for training which is a difficult task for large datasets as they need manual annotation. The text-based strategy also has an additional issue of audio being produced at the end of each word which produces a delay in speech generation. As a result, it becomes inappropriate for real-time applications. Unnatural speech is produced by text-based strategy due to the inability to capture intonation and emotion. In the proposed methodology direct silent video-to-speech conversion is done since audio samples are generated for each video sequence using regression which results in natural speech. Most videos provide a corresponding audio track, so model training is done under self-supervised conditions. Thus, our method solves the limitation of the text-based approach.

Certain shortcomings were identified in the proposed approach after investigation, and further development will be pursued to achieve better performance. Potential future research directions include:

1. Utilizing features from the entire facial region, rather than just the mouth, to reconstruct speech signals.
2. Implementing a model capable of withstanding external factors such as background noise, lighting conditions, and other environmental influences.
3. Developing a model that incorporates emotions into the generated voice signal.
4. Creating a model that requires fewer training parameters.
5. Constructing a model that supports multiple languages.
6. Enhancing the alignment between speech and lip movements to produce more accurate speech signals.
7. Integrating optimization algorithms to find the most effective solutions for improving overall performance.

A complete list of abbreviations is listed in *Appendix I.*

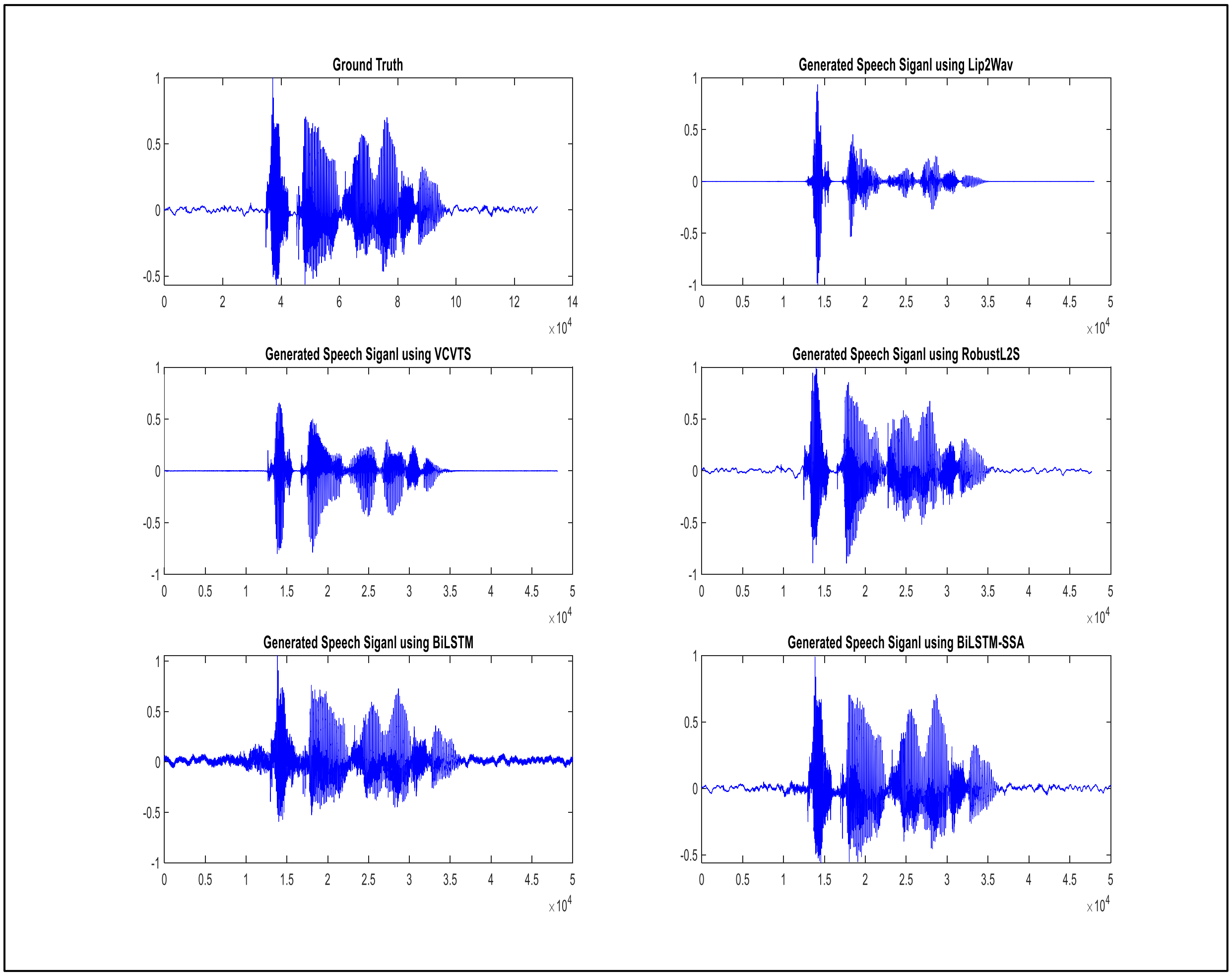


Figure 10 Comparison of synthesized speech for male speaker-1 with state-of-the-art lip-to-speech synthesis

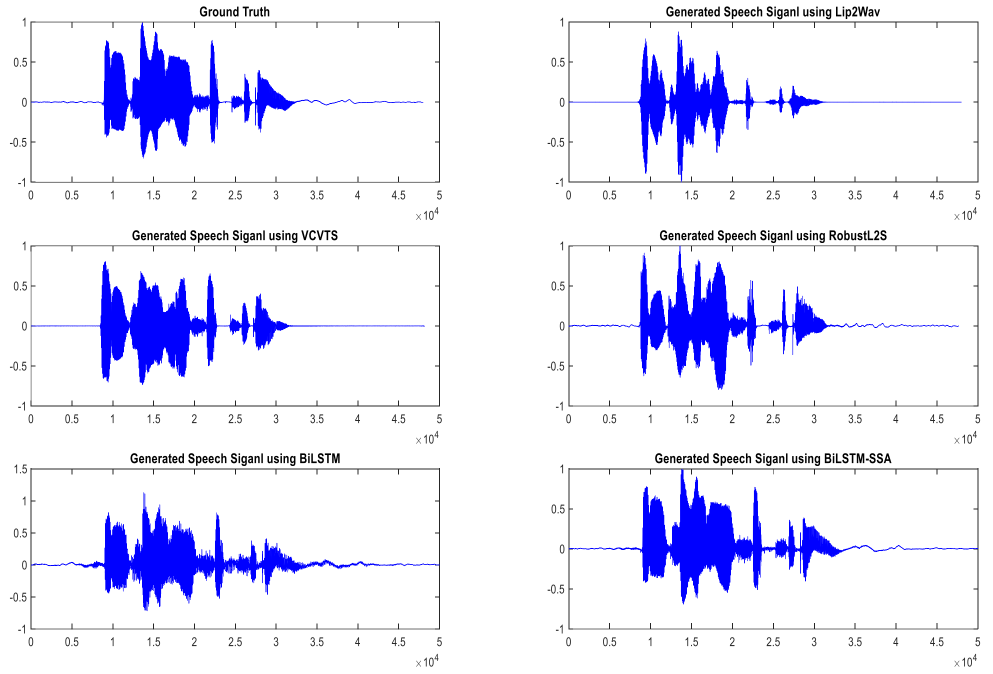


Figure 11 Comparison of synthesized speech for female speaker-4 with state-of-the-art lip-to-speech synthesis

6.Conclusion and future work

The outcomes were achieved by introducing the two suggested architectures-BILSTM without optimization and SSA optimization with BILSTM. In the proposed technique, key frames are selected at intervals of three, which helps to reduce computation and improve execution speed. The performance parameters like MSE and RMSE are reduced for both models during the training phase. The speech quality parameters SNR, STOI, and PESQ using SSA optimization show slight improvement compared to the BiLSTM without optimization. The proposed model SSA optimization with BiLSTM for the reconstruction of the speech signal is compared with several deep learning models. It shows the improvement in STOI and PESQ for unseen speakers of the GRID dataset. PESQ parameter is significantly improved by 4.06 and 0.202 improvement in STOI over VCA-GAN. The visual features are extracted from ResNet-50 which is deep in architecture. It is observed that a deeper convolution layer can find out the accurate shape of the object and thus improve the results.

The work described in this paper can be improved upon by increasing the intelligibility of speech reconstruction from an unconstrained dictionary and a more realistic configuration. The proposed model can be used as a basis for different speech-oriented applications such as speech enhancement when it is corrupted. We can extend the work to separate the mixed speech signal from the noisy environment. The suggested system is useful to listeners who have trouble understanding speakers with foreign accents. The lip movements of foreign speakers for producing the standard alphabet are the same. Using the proposed methodology, the video sequence from the foreign speaker would stay the same but the audio could be changed. We can further update the architecture and information exchange can be made more effective by substituting familiar audio for the audio used to train the model. This way, even when the words are spoken by foreign speakers, the listener will recognize the sound. The future work can be extended for educational purposes such as massively open online courses (MOOC). The audience is left wondering what the speaker said whenever the speech is not auditable. The feed from multiple cameras can easily be used to reconstruct the speech and sound of the speaker.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The GRID audio-visual dataset utilized in the experimentation is publicly accessible and can be found at https://spandh.dcs.shef.ac.uk/gridcorpus/.

Author's contribution statement

Shilpa Sonawane: Conceptualization, literature review, data collection, design, implementation, writing- original draft and editing, analysis and interpretation of results. **P. Malathi:** Study conception, supervision, investigation on challenges and draft reviewing.

References

[1] Devi S, Chokshi S, Kotian K, Warwatkar J. Visual speech recognition. In 4th Biennial international conference on nascent technologies in engineering 2021 (pp. 1-4). IEEE.

[2] Gabbay A, Ephrat A, Halperin T, Peleg S. Seeing through noise: visually driven speaker separation and enhancement. In international conference on acoustics, speech and signal processing 2018 (pp. 3051-5). IEEE.

[3] Stewart D, Seymour R, Pass A, Ming J. Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Transactions on Cybernetics. 2013; 44(2):175-84.

[4] Lesani FS, Ghazvini FF, Dianat R. Mobile phone security using automatic lip reading. In 9th international conference on e-commerce in developing countries: with focus on e-business 2015 (pp. 1-5). IEEE.

[5] Mathulaprangsan S, Wang CY, Kusum AZ, Tai TC, Wang JC. A survey of visual lip reading and lip-password verification. In international conference on orange technologies 2015 (pp. 22-5). IEEE.

[6] Sengupta S, Bhattacharya A, Desai P, Gupta A. Automated lip reading technique for password authentication. International Journal of Applied Information Systems. 2012; 4(3):18-24.

[7] Son CJ, Senior A, Vinyals O, Zisserman A. Lip reading sentences in the wild. In proceedings of the conference on computer vision and pattern recognition 2017 (pp. 3444-53). IEEE.

[8] Ephrat A, Halperin T, Peleg S. Improved speech reconstruction from silent video. In proceedings of the international conference on computer vision workshops 2017 (pp. 455-62). IEEE.

[9] Liu J, Li C, Ren Y, Chen F, Zhao Z. Diffsinger: singing voice synthesis via shallow diffusion mechanism. In proceedings of the conference on artificial intelligence 2022 (pp. 11020-8). AAAI.

[10] Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. PLoS Computational Biology. 2016; 12(11):e1005119.

[11] Gabbay A, Shamir A, Peleg S. Visual speech enhancement. In proceedings of the workshop on interspeech 2018 (pp. 1170-4).

[12] Mattos AB, Oliveira DA. Multi-view mouth renderization for assisting lip-reading. In proceedings of the 15th international web for all conference 2018 (pp. 1-10). ACM.

[13] Deshmukh N, Ahire A, Bhandari SH, Mali A, Warkari K. Vision based lip reading system using deep learning. In international conference on computing, communication and green engineering 2021 (pp. 1-6). IEEE.

[14] Ali NH, Abdulmunem ME, Ali AE. Constructed model for micro-content recognition in lip reading based deep learning. Bulletin of Electrical Engineering and Informatics. 2021; 10(5):2557-65.

[15] Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R. Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 24(2):198-213.

[16] Cox SJ, Harvey RW, Lan Y, Newman JL, Theobald BJ. The challenge of multispeaker lip-reading. In AVSP 2008 (pp. 179-84).

[17] Ortega A, Sukno F, Lleida E, Frangi AF, Miguel A, Buera L, et al. AVCAR: a Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In LREC 2004:763-6.

[18] Mahboob K, Nizami H, Ali F, Alvi F. Sentences prediction based on automatic lip-reading detection with deep learning convolutional neural networks using video-based features. In soft computing in data science: 6th international conference, virtual event, proceedings 2021 (pp. 42-53). Springer Singapore.

[19] Caranica A, Cucu H, Burileanu C, Portet F, Vacher M. Speech recognition results for voice-controlled assistive applications. In international conference on speech technology and human-computer dialogue 2017 (pp. 1-8). IEEE.

[20] Kumar K, Chen T, Stern RM. Profile view lip reading. In international conference on acoustics, speech and signal processing 2007 (pp. 429-32). IEEE.

[21] Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018; 44(12):8717-27.

[22] Fernandez-lopez A, Sukno FM. Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing. 2018; 78:53-72.

[23] Ivanko D, Ryumin D, Karpov A. Automatic lip-reading of hearing impaired people. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2019; 42:97-101.

[24] Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Audio-visual speech recognition using deep learning. Applied Intelligence. 2015; 42:722-37.

[25] Thangthai K, Harvey R. Improving computer lipreading via DNN sequence discriminative training techniques. In proceedings 2017 (pp. 3657-61). ISCA.

[26] Le CT, Milner B. Reconstructing intelligible audio speech from visual speech features. In interspeech 2015 (pp. 3355-9). ISCA.

[27] Rekik A, Ben-hamadou A, Mahdi W. An adaptive approach for lip-reading using image and depth data.

Multimedia Tools and Applications. 2016; 75:8609-36.

[28] Kumar Y, Jain R, Salik KM, Shah RR, Yin Y, Zimmermann R. Lipper: synthesizing thy speech using multi-view lipreading. In proceedings of the AAAI conference on artificial intelligence 2019 (pp. 2588-95). AAAI.

[29] Prajwal KR, Mukhopadhyay R, Namboodiri VP, Jawahar CV. Learning individual speaking styles for accurate lip to speech synthesis. In proceedings of the conference on computer vision and pattern recognition 2020 (pp. 13793-802). IEEE.

[30] Akbari H, Arora H, Cao L, Mesgarani N. Lip2audspec: speech reconstruction from silent lip movements video. In international conference on acoustics, speech and signal processing (ICASSP) 2018 (pp. 2516-20). IEEE.

[31] Kumar LA, Renuka DK, Rose SL, Wartana IM. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. International Journal of Cognitive Computing in Engineering. 2022; 3:24-30.

[32] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. In proceedings of interspeech 2017 (pp. 3652-6).

[33] Kim M, Hong J, Park SJ, Ro YM. Cromm-vsr: cross-modal memory augmented visual speech recognition. IEEE Transactions on Multimedia. 2021; 24:4342-55.

[34] Yang Q, Bai Y, Liu F, Zhang W. Integrated visual transformer and flash attention for lip-to-speech generation GAN. Scientific Reports. 2024; 14(1):1-12.

[35] Qu L, Weber C, Wermter S. Lipsound2: self-supervised pre-training for lip-to-speech reconstruction and lip reading. IEEE Transactions on Neural Networks and Learning Systems. 2022; 35(2):2772-82.

[36] Kim M, Yeo JH, Choi J, Ro YM. Lip reading for low-resource languages by learning and combining general speech knowledge and language-specific knowledge. In proceedings of the IEEE/CVF international conference on computer vision 2023 (pp. 15313-25). IEEE.

[37] Weng Z, Qin Z, Tao X, Pan C, Liu G, Li GY. Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Transactions on Wireless Communications. 2023; 22(9):6227-40.

[38] Ivanko D, Ryumina E, Ryumin D. Improved automatic lip-reading based on the evaluation of intensity level of speaker's emotion. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2023; 48:89-94.

[39] Hegde S, Mukhopadhyay R, Jawahar CV, Namboodiri V. Towards accurate lip-to-speech synthesis in-the-wild. In proceedings of the 31st international conference on multimedia 2023 (pp. 5523-31). ACM.

[40] Yemini Y, Shamsian A, Bracha L, Gannot S, Fetaya E. LipVoicer: generating speech from silent videos guided by lip reading. In the twelfth international conference on learning representations 2024 (pp.1-20).

[41] Cooke M, Barker J, Cunningham S, Xu S. The grid audio-visual speech corpus (1.0). Zenodo: Geneva, Switzerland. 2006.

[42] King DE. Dlib-ml: a machine learning toolkit. The Journal of Machine Learning Research. 2009; 10:1755-8.

[43] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-8). IEEE.

[44] Ouyang C, Zhu D, Wang F. A learning sparrow search algorithm. Computational Intelligence and Neuroscience. 2021; 2021(1):1-23.

[45] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In international conference on acoustics, speech, and signal processing 2001(pp. 749-52). IEEE.

[46] Mira R, Vougioukas K, Ma P, Petridis S, Schuller BW, Pantic M. End-to-end video-to-speech synthesis using generative adversarial networks. IEEE Transactions on Cybernetics. 2022; 53(6):3454-66.

[47] Mira R, Haliassos A, Petridis S, Schuller BW, Pantic M. SVTS: scalable video-to-speech synthesis. In proceedings of the interspeech 2022 (pp. 1836-40).

[48] Kim M, Hong J, Ro YM. Lip to speech synthesis with visual context attentional GAN. Advances in Neural Information Processing Systems. 2021; 34:2758-70.

[49] Sahipjohn N, Shah N, Tambrahalli V, Gandhi V. RobustL2S: speaker-specific lip-to-speech synthesis exploiting self-supervised representations. In Asia pacific signal and information processing association annual summit and conference 2023 (pp. 1492-9). IEEE.

[50] Wang D, Yang S, Su D, Liu X, Yu D, Meng H. VCVTS: multi-speaker video-to-speech synthesis via cross-modal knowledge transfer from voice conversion. In international conference on acoustics, speech and signal processing 2022 (pp. 7252-6). IEEE.

Shilpa Sonawane received her ME degree in Digital Systems from Pune University in 2012 and is currently pursuing a Ph.D. at the D.Y. Patil College of Engineering Research Centre, affiliated with Savitribai Phule Pune University. Her research areas include Signal Processing, Speech Processing, Image Processing, Machine Learning, and Deep Learning.
Email: shilpa.sonawane8@gmail.com

P. Malathi received her ME degree in Digital Systems from Pune University in 2001 and her Ph.D. in Wireless Communication from Pune University in 2011. She is currently serving as a Professor and Principal at D.Y. Patil College of Engineering, Pune. Her research areas include the design of Microstrip Antennas, Patch Antennas, Monopole Antennas, and Wheel-Shaped Fractal Antennas using Artificial Neural Networks, as well as Signal Processing, Image Processing, VLSI, and Embedded Systems.
Email: pjmalathi@dypcoeakurdi.ac.in

Appendix I

S. No.	Abbreviation	Description
1	ASR	Automatic Speech Recognition
2	AAM	Active Appearance Model
3	AVSR	Audio Visual Speech Recognition
4	AWGN	Additive White Gaussian Noise
5	BERT	Bidirectional Encoder Representations from Transformers
6	BiLSTM	Bidirectional Long Short-Term Memory
7	CER	Character Error Rate
8	CN-CVS	Chinese Continuous Visual Speech
9	CNN	Convolutional Neural Network
10	DCT	Discrete Cosine Transform
11	DNN	Deep Neural Network
12	ESTOI	Extended Short-Time Objective Intelligibility
13	FA-GAN	Flash Attention GAN
14	FC	Fully Connected
15	GAN	Generative Adversarial Network
16	GMM	Gaussian Mixture Model
17	GRU	Gated Recurrent Unit
18	HMM	Hidden Markov Model
19	HOG	Histogram of Oriented Gradients
20	ITU-T	International Telecommunication Union Telecommunication Standardization Sector
21	KNN	K-Nearest Neighbour
22	LM	Language Memory
23	LPC	Linear Predictive Coefficient
24	L2S	Lip-to-Speech
25	LSTM	Long Short-Term Memory
26	MFCC	Mel Frequency Cepstral Coefficients
27	MSE	Mean Squared Error
28	PCA	Principal Component Analysis
29	PESQ	Perceptual Evaluation of Speech Quality
30	ReLU	Rectified Linear Unit
31	ResNet-50	Residual Neural Network-50
32	RMSE	Root Mean Square Error
33	ROI	Region of Interest
34	SNR	Signal to Noise Ratio
35	SSA	Sparrow Search Algorithm
36	STOI	Short Time Objective Intelligibility
37	SVM	Support Vector Machine
38	TTS	Text-To-Speech
39	VAM	Visual Audio Memory
40	VGG	Visual Geometry Group19
41	VSR	Visual Speech Recognition
42	VSS	Visual Speech Synthesis
43	VCVTS	Voice Conversion-based Video-To-Speech
44	WER	Word Error Rate
45	3-D	Three-Dimensional
46	VCA-GAN	Visual Context Attentional-Generative Adversarial Network
47	MOOC	Massively Open Online Courses