



Combining audio and visual speech recognition using LSTM and deep convolutional neural network

R. Shashidhar¹ · S. Patilkulkarni¹ · S. B. Puneeth²

Received: 17 November 2021 / Accepted: 13 February 2022 / Published online: 24 February 2022

© The Author(s), under exclusive licence to Bharati Vidyapeeth's Institute of Computer Applications and Management 2022

Abstract Human speech is bimodal, whereas audio speech relates to the speaker's acoustic waveform. Lip motions are referred to as visual speech. Audiovisual Speech Recognition is one of the emerging fields of research, particularly when audio is corrupted by noise. In the proposed AVSR system, a custom dataset was designed for English Language. Mel Frequency Cepstral Coefficients technique was used for audio processing and the Long Short-Term Memory (LSTM) method for visual speech recognition. Finally, integrate the audio and visual into a single platform using a deep neural network. From the result, it was evident that the accuracy was 90% for audio speech recognition, 71% for visual speech recognition, and 91% for audiovisual speech recognition, the result was better than the existing approaches. Ultimately model was skilled at enchanting many suitable decisions while forecasting the spoken word for the dataset that was used.

Keywords Audio-visual speech recognition · Lip-reading · DNN · LSTM · Custom Dataset

1 Introduction

Audiovisual speech recognition is the most widely used technique today to automatically detect what a person is saying in the form of text. In the modern era, it has gained a lot of popularity and we use it almost in our day-to-day life in the form of Google Assistant or even Amazon Alexa. However, the common observation is that this audio speech recognition is mostly used indoors and does not give a good response outdoors. This is due to an intervention of noise. The noise adds to the audio signals and most of the necessary data are lost. That is not the case when it comes to Visual Speech Recognition (VSR), which has advantages over Audio Speech Recognition. They are (a) it is not attentive to audio noise and modification in audio environments has no effect on the data. (b) Does not need the user to make a sound. In present times we have a lot of data available and even possess a high computational ability.

AVSR primarily consists of two main part; The Audio recognition and Visual recognition (Lip reading). While, the audio recognition consists of feature extraction and recognition processes, the video recognition consists of face detection, lip localization, feature extraction and recognition. Combined, these two sources of speech information result in better automatic recognition rates than were obtained from either source alone. We chose to map the visual signal into an acoustic representation closely related to the vocal tract's transfer function. Given such a mapping, the visual signal could be converted and then integrated with the acoustic signal prior to any symbolic encoding. The objectives of the work is given below.

- (a) Develop a database for English Language.
- (b) Audio feature extraction using MFCC and Classification using 1D CNN.

✉ R. Shashidhar
shashidhar.r@sjce.ac.in

S. Patilkulkarni
sudarshan_pk@sjce.ac.in

S. B. Puneeth
puneeth.sb@presidencyuniversity.in

¹ Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysore 570006, India

² Departments of Electronics and Communication Engineering, Presidency University, Bangalore 560064, India

- (c) Develop an algorithm for lip localization.
- (d) Develop an LSTM algorithm for Visual speech recognition.
- (e) Integration of Audio and Visual Speech using Deep Neural Network.
- (f) Comparison of Proposed Result with Existing Results.

The rest of this paper is scheduled as follows: In Sect. 2 discussed the literature review of the existing audio visual speech recognition methods. In section discussed the database details. In Sect. 4 explained the proposed methodology. In Sect. 5 discuss the result and discussion of the proposed method. In Sect. 6 concludes the proposed work.

2 Literature review

An extensive literature survey has been conducted prior to the beginning of the proposed work. In this section discussed the existing algorithms used for audiovisual speech recognition and also figure out the drawback of the existing systems.

This will also serve us with an advantage in using various machine learning and deep learning algorithms to get the best results possible. The LRS2 database is used as the most common database available [1]. The feature extraction is carried out within the Region of Interest. The performance observed by the audio speech recognition was not on par compared to the performance given by the AVSR in noisy conditions. The noise can be of different types like street, train, etc. It shows that the noise independent of the type gives the same results. Lip-reading is the job of deciphering a transcript from the measure of a presenter's mouth. Ahmad B. A. Hassanat explained different approaches to lip localization [2]. Ayaz A. Shaikh et al. proposed the depth sensor camera has also been used to get the third dimension in the dataset. During the creation of the dataset, the above-mentioned factors have been taken care of by using a headrest [3]. Themis Stafylakis et al. proposed residual and LSTM techniques for the LRW database and get 83% accuracy [4]. Shillingford et al. have used Lipnet. In the lipnet, used two approaches to solve the problems are learning the visual features and prediction. The word error rate of this work is 89.8% and 76.8% [5].

One of the other architectures to implement Lip-reading is Long-Short Term Memory (LSTM) [6]. LSTM is used for lip-reading that determines the words from the video input, it is accomplished by selectively indicating spatiotemporal balances that are important for an individual dataset. LRS2 datasets were used in the model and it achieves 85.2%. G. Sterpu et al. [7] look into futuristic Deep

Neural Network architectures for lip-reading founded on a sequence-to-sequence Recurrent Neural Network. This work makes sure for both redeveloped and 2D or 3D Convolutional Neural network visual frontends, operational monotonic consideration, and a combined connectionist Temporal Classification Sequence-to-sequence loss. This evaluated system is done with fifty-nine talkers and terminology of over six thousand arguments on the widely accessible TCD-TIMIT dataset. Kumar et al. [8] showed the set of experiments in detail for speaker-dependent, out-of-vocabulary, and speaker-independent settings. To show the real-time nature of audio produced in the system, the hindrance values of Lipper have been compared with other speech reading systems. The audio-only accuracy is 80.25%, the annotation accuracy variance is 2.72% in audio, and Audio-visual accuracy is 81.25%, the annotation accuracy variance is 1.97% in audio-visual. One of the common datasets used in lip reading is Grid audio-visual dataset, the work in [9] is based on the Grid audio-visual dataset. The visual dataset is recorded with a frame rate is 25 Frames per second, a total of 75 frames per sample for 3 s.

In this work, LCA Net and end-to-end deep neural networks. The system archives 1.3% CER and the word error rate is 3.0%. Dilip Kumar et al. suggested the new-fangled SD-2D-CNN-BLSTM [10] architecture. The analysis of two different approaches like 3D-2D-Convolutional neural network-BLSTM trained with CTC loss on Characters and 3D-2D-Convolutional neural network- Bidirectional Long Short-Term Memory (CNN-BLSTM) trained with CTC loss on word labels for lip-reading is presented. For the first approach, the word error rate is 3.2% and 15.2% for seen and unseen words respectively. Performance on-grid dataset of the second approach, the word error rate is 1.3% and 8.6% for seen and unseen words respectively. The performance of the Indian English unseen dataset word error rate is 19.6% and 12.3% for the two approaches. One of the most famous datasets used for lip reading is "Lip Reading in the wild (LRW)" [11] from BBC Tv it contains 500 targeted words. Themis Stafylakis et al. used Residual networks and Bidirectional LSTMs and the misclassification rate of the architecture is 11.92%. Using the same database and the same method got 83% accuracy.

Audiovisual speech recognition is one prospective explanation for speech recognition in a noisy environment [12]. Shiliang Zhang et al. used bimodal -DFNN, used 150 h of multi-condition training data, and archives a 12.6% phone error rate for clean test data. The Word error rate is 29.98%. Kuniaki Noda et al. introduce a multi-stream HMM model for integration of Audio and Visual features [13]. The Word Recognition rate of MSHMM is 65% and the Signal noise ratio is 10 dB. Stavros Petridis

et al. Long–short Memory based end to end visual speech recognition classification [14]. The model contains two flows which citation features straight away from the mouth. The two streams take place via bidirectional Long Short Term Memory. Databases like ouluVS2 and CUAVE used, the accuracy of the work is 9.7% and 1.5% respectively. Fei Tao et al. [15] proposed structure is likened with Conventional Hidden Marko Model with observation models fulfilled with Gaussian mixture model and used this channel matched word error rate is 3.70% and Channel mismatched word error rate is 11.48%. The hybrid Connectionist Temporal Classification architecture for audiovisual recognition of speech in the wild is used in the [16]. The audio features are of many kinds. The three of them used in [17] are LPC, PLP, and MFCC.

The study shows that the MFCC has the highest accuracy of about 94.6% for the Hindi Language in a noiseless environment. It proceeds a lot of periods to create and process the data to be in the format required for the application. The objective that is defined in the work [18] can be affected by the varying light intensity, movement of the head, the distance from the camera. Ochiai et al. proposed the most significant speaker clues are extracted from the dataset. This is attention-based feature extraction. They have used 3 layers of BLSTM with 512 units each [19]. Joon Son Chung et al. proposed a new set of databases called LRS it contains 100,000 normal sentences from BBC television [20]. Namboodiri et al. used Charlie Chaplin videos; the word spotting technique achieves 35% upper despicable typical accuracy over recognition-based technique on extensive LRW dataset. Determine the request of the technique by word recognizing in a standard speech video are “The great dictator” by Charlie Chaplin [21]. Thabet et al. applied machine learning methods to identify lip interpretation and three classifiers became the preminent outcomes which are Gradient Boosting, Support vector machine, and logistic regression with outcomes 64.7%, 63.5%, and 59.4% correspondingly [22].

Yaman Kumar et al. proposed a speech reading or lip-reading is the method of empathizing and receiving phonetic topographies from a presenter’s visual features such as movement of mouths, face, teeth, and tongue [23]. Lu et al. proposed technology for visual speech recognition which association’s machine visualization and linguistic perception [24].

Iain et al. created their custom database called AV letters and they used three approaches first one is the hidden Marko model for recognition, for lip features top-down method is used and the third one is a bottom-up method used for nonlinear scale analysis [25]. Abderrahim Mesbah et al. proposed Hahn CNN for three databases like AV letters; Oulu VS2 and BBC LRW got an accuracy of 59.23%, 93.72%, and 58.02% [26]. Shashidhar et al.

proposed the VGG16 CNN method for visual speech recognition and in this experiment; they used custom datasets and got 76% accuracy [27]. Ximeng et al. proposed the multi-layer feature fusion convolution neural network for audiovisual speech recognition and apply MFFCNN to TCD-TIMIT and GRID Corpus dataset and got an accuracy of 82.7% [28]. Weijiang Feng et al. proposed the Multimodal recurrent neural network for audiovisual speech recognition and MRNN is applied to the AV Letter dataset and got an accuracy of 84.4% [29].

3 Database

In this section discuss the dataset creation steps, dataset features and also discuss the challenges are faced when we creating the dataset for English Language.

3.1 Dataset creation

Data-set is created for both English Words using an extensive setup which includes an electronic gimbal for stable video and a Smartphone with sufficient storage space. In Table 1 mention the parameter of the dataset features. The dataset is embraced of interrelated audio and lip movement data in various videos of multiple topics construing identical words. The formation of the dataset was finished to enable the progress and proof of procedures charity to train and test the method that contains lip-motion. The data set is a gathering of videos of agrees declaiming a fixed screenplay that is planned to be used to train software to recognize lip-motion patterns.

The recordings were collected in a controlled, noise-free, indoor setting with a smartphone capable of recording at 4 K resolution. This data set consists of around 240 video samples per person. 11 male and 13 female subjects, with ages ranging from 18 to 30, volunteered for the dataset creation process. This data set can be used for speech recognition, lip reading applications. Around 240 video samples were collected per subject.

Table 1 Dataset features

Parameter	Value
Resolution	1080 × 1920 P
Frame/Second	60 FPS
Storage Duration of Video	1.20 PS
Storage Size of video	10 MB

3.2 Challenges while creating dataset

Various challenges were encountered during the data-set creation process which is explained below.

- Interference of external noise may disrupt audio feature extraction. A noise-free environment is an important requirement of data-set creation.
- Lip movement of an individual should be in conjunction with each other to extract the lip feature, random movement of lip leads to error.
- Each person who is ready to give a database has to spare around 30–45 min reciting the words, which can be tedious.
- Recording a video of a person with a mustache or beard leads to difficulty in detecting lip movement.
- The selection of English and Kannada words to prepare the database was difficult as some of the words have similar pronunciations.

4 Methodology

This section deals with the pipeline and the methodologies that were implemented in AVSR. To implement AVSR, the custom database was created and the language that was used in the database was English, which comprised of seven different words. The seven words are ‘About’, ‘Bottle’, ‘Dog’, ‘English’, ‘Good’, ‘People’, ‘Today’. Fifteen individuals pronounced each word five times, and it was recorded. Therefore, in total (15persons \times 7 words \times 5 times = 525 videos) the dataset consists of five hundred twenty-five videos. Out of these, 420 videos were used for training the model and the remaining were used for testing.

4.1 Audio speech model

First audio files are created from the video dataset and saved in.wav formats using FFmpeg. Then the features are extracted from the audio using Libros which is an open-source module that is available in python. The five features that are extracted from audio are MFCCS, CHROMA, MEL, CONTRAST, and TONNETZ. All these features are combined to get generate the feature vector of size 193×1 . Next, a Convolutional Neural Network using 1 Conv1D layer, followed by MaxPooling1D layer, Batch normalization layer, and dropout and then followed by two Dense Layers is created. One-dimensional CNNs work with a sequence in one dimension, and tends to be useful in various signal analysis over fixed-length signals. They work well for the analysis of audio signals, for instance. The

output from the corresponding layers of the audio model will be as follow.

$$a^1 + b^1 = \sum_{i=1}^{193} \text{conv1D}(w_i, X_{\text{tran_audio}}[i]) \quad (1)$$

$$y^1 = R(a^1) \quad (2)$$

$$y^2 = R(w^2 y^1 + b^2) \quad (3)$$

$$y^3 = R(w^3 y^2 + b^3) \quad (4)$$

$$y^4 = R(w^4 y^3 + b^4) \quad (5)$$

where y^i is the output vector of layer i , R is the ReLu activation function, w^i is the weights of layer i , b^i is the bias of layer i . Finally, a SoftMax layer was attached to this network for the classification. The loss function which is used to train the model is cross-entropy which is given by

$$L^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log(\hat{y}^{<t>}) - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>}) \quad (6)$$

4.2 Visual model

First mouth region is extracted from the video using dlib library which is available in python 3 as shown in Fig. 1 in each frame of the video. Then the regions are converted into grey scale to reduce the complexity of the model as shown in Fig. 2 Then the position of the outer lip coordinates are extracted and saved in the feature vector.

Recognition of visual speech using Long Short-Term Memory (LSTM), Fig. 3 shows the structure of LSTM Cell. Then a model with a network of LSTM's and dense layers (Deep LSTM Network) is created. Long Short-Term Memory (LSTM) network is a type of recurrent neural network which is can learn order dependence in a sequence

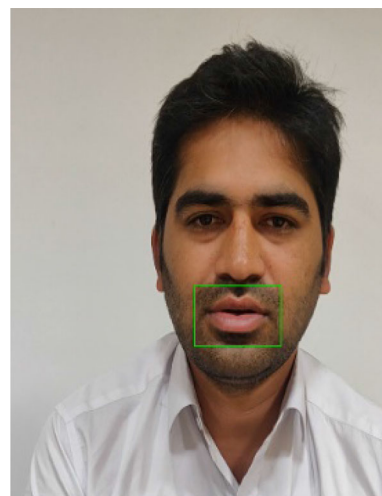


Fig. 1 Mouth ROI Extraction



Fig. 2 Conversion to gray scale

prediction problem. It contains three gates: the first one is input gate, forget gate, and output gate. The model which is created contains an LSTM layer with 8 hidden units and 30-time stamps as the first layer. The operations inside the LSTM cell are as follows. Forget Gate decides whether to keep or forget the info from the previous timestamps and Input Gate quantifies the importance of that data coming as an input and the Output Gate figures the most relevant output that it must generate.

While using LSTM tanh activation function is used, the structures of the tanh activation function as shown in Fig. 4.

Tanh means hyperbolic tangent function it is like a sigmoid activation function.

The function accepts any real value as input and returns a value between -1 and 1 . The larger the input means more positive values, the closer the output to 1.0 , and the smaller the input means more negative values, the closer the output to -1 .

Fig. 3 Structure of an LSTM cell

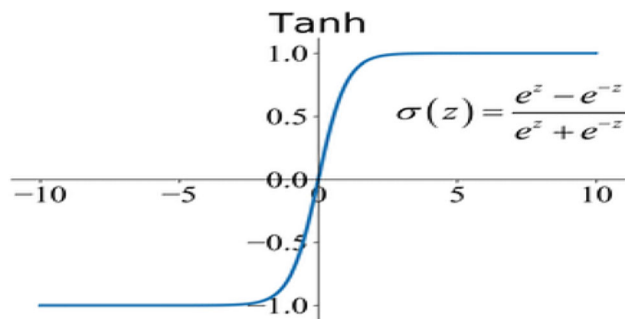
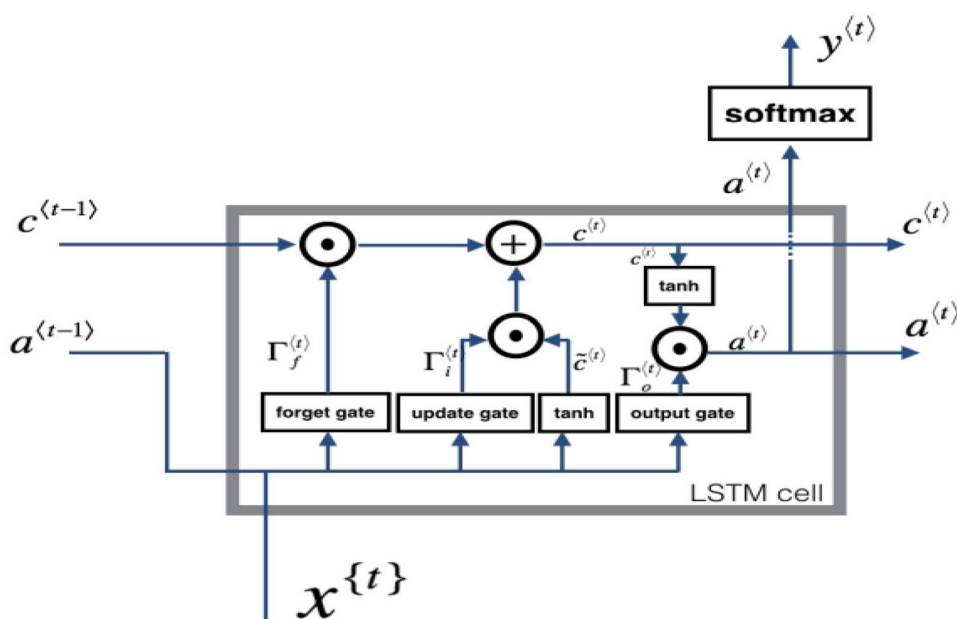


Fig. 4 Performance of Tanh Activation Function

$$\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{7}$$

The model which is created contains an LSTM layer with 8 hidden units and 30-time stamps as the first layer. The operations inside the LSTM cell are as follows.

$$\tilde{c}\langle t \rangle = \tanh (W_c[a\langle t - 1 \rangle, x\langle t \rangle] + b_c) \tag{8}$$

$$\Gamma_u = \sigma(W_u[a\langle t - 1 \rangle, x\langle t \rangle] + b_u) \tag{9}$$

$$\Gamma_f = \sigma(W_f[a\langle t - 1 \rangle, x\langle t \rangle] + b_f) \tag{10}$$

$$\Gamma_o = \sigma(W_o[a\langle t - 1 \rangle, x\langle t \rangle] + b_o) \tag{11}$$

$$c\langle t \rangle = \Gamma_u * \tilde{c}\langle t \rangle + \Gamma_f * c\langle t - 1 \rangle \tag{12}$$

$$a\langle t \rangle = \Gamma_o * \tanh(c\langle t \rangle) \tag{13}$$

where c represents the memory cell, t represents the time stamp, Γ_u represents the update gate. Γ_f represents the forget gate, Γ_o represents the output gate, σ represents the sigmoid function, W_u represents the weights of update gate, W_f represents the weights of forget gate, W_o

represents the weights of output gate, b represents bias, c represents candidate cell variables.

Next, one more LSTM layer is introduced. So the resulting output from the second layer will be

$$\tilde{c}1^{(t)} = \tanh(W1_c[a1^{(t-1)}, a^{(t)}] + b1_c) \tag{14}$$

$$\Gamma1_f = \sigma(W1_f[a1^{(t-1)}, x1^{(t)}] + b1_f) \tag{15}$$

$$\Gamma1_o = \sigma(W1_o[a1^{(t-1)}, x1^{(t)}] + b1_o) \tag{16}$$

$$c^{(t)} = \Gamma1_u * \tilde{c}^{(t)} + \Gamma1_f * c1^{(t-1)} \tag{17}$$

$$a^{(t)} = \Gamma1_o * \tanh(c1^{(t)}) \tag{18}$$

This is followed by three dense layers. Hence, again the output equations from these three dense layers will be.

$$y^2 = R(W2 * a1 + b2) \tag{19}$$

$$y^3 = R(W3 * y^2 + b3) \tag{20}$$

$$y^4 = R(W4 * y^3 + b4) \tag{21}$$

and finally, a softmax layer was attached for to this network for the classification. The loss function which is used to train the model is cross entropy which is given by

$$L^{(t)}(y^{(t)}, \hat{y}^{(t)}) = -y^{(t)} \log(\hat{y}^{(t)}) - (1 - y^{(t)}) \log(1 - \hat{y}^{(t)}) \tag{22}$$

4.3 Fusion model

Integration of Audiovisual speech recognition using a deep neural network. Integration of audiovisual contains three parts, the first one is the audio-only second one is visual only third one is a integration of the audio and visual. Figure 5 shows the deep convolutional neural network, in this model one input layer, one output layer and two hidden layers are used.

In the audio-only part, features are extracted in the same way as the Audio model. Then a deep Convolutional Neural Network is created. The model which is created is

the replica of the Audio model except in place of a softmax layer there is an additional dense layer.

$$y^5 = R(W^5 y^4 + b^5) \tag{23}$$

In the Video-only part, Video features as extracted the same way as the Video model. Then a deep LSTM network is created. The first layer which is the LSTM layer contains 128 hidden units with 8-time stamps. This LSTM layer is followed by a dropout and a dense layer. So, the resulting equation from this dense layer will be,

$$y_v = R(W2 * a1^{(t)} + b2) \tag{24}$$

In ‘‘Combination of Audio-only and Video-only parts’’ the feature map from the first dense layer from the Audio-only part is concatenated with the feature map from the first LSTM layer from Video only part. From Eqs. (3) and (13),

$$a_c = [a1^{(t)}, y^2] \tag{25}$$

The resulting feature map is passed on to a deep neural network which contains three dense layers while the first two dense layers are followed by a Batch normalization layer and a dropout layer respectively

$$y_{d1} = R(W_{d1} * a_o + b_{d1}) \tag{26}$$

$$y_{d2} = R(W_{d2} * y_{d1} + b_{d2}) \tag{27}$$

$$y_{d3} = R(W_{d3} * y_{d2} + b_{d3}) \tag{28}$$

And as the final step, all the above three parts are combined so the vector formed by this will be a combination of the output vector of all the above three parts. Therefore, from Eqs. (23), (24) and (25)

$$ac_2 = [y^5, y_v, y_{d3}] \tag{29}$$

Then this is passed on to a deep neural network which contains three dense layers followed b a batch normalization layer and a dropout layer.

$$y_{c1} = R(W_{c1} * ac_2 + b_{c1}) \tag{30}$$

$$y_{c2} = R(W_{c2} * y_{c1} + b_{c2}) \tag{31}$$

$$y_{c3} = R(W_{c3} * y_{c2} + b_{c3}) \tag{32}$$

5 Result and discussion

This section discusses the result of the audio-only model, visual-only model, and audio-visual model with accuracy curve, loss curve, and confusion matrix and with classification table.

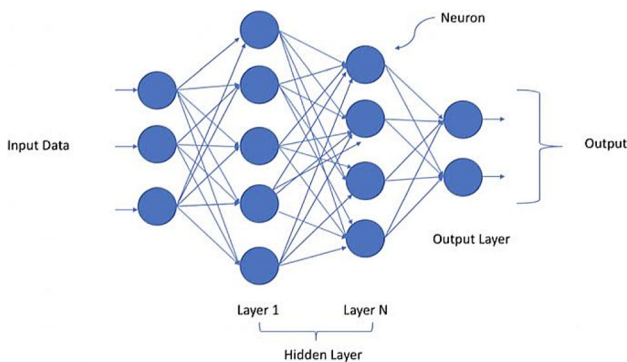


Fig. 5 Deep convolutional neural networks

5.1 Audio speech recognition evaluation

Figure 6 shows the training of epoch for audio-only model and here shows the training accuracy 90.48% and testing accuracy of 96.62%. Figure 7 shows the accuracy curve of the audio-only model.

Figure 8 shows the loss curve of the audio-only model. Figure 9 shows the accuracy and misclassification of each word. The first word “About” is being recognized with 80% accuracy and misclassification is only 20%, which means the algorithm predicted 80% as “About” and 20% as “Today” as in the graph. The second word is “Bottle”. It is recognized with 86% accuracy and misclassification is 14%, which means the algorithm predicted 86% as “Bottle” and 14% as “Dog” as in the graph. The third word “Dog” is recognized with 86% accuracy and misclassification is 14%, which means it predicted 86% as “Dog” and 14% as “About”, as shown in the graph. The fourth word “English” is recognized with 100% accuracy and no misclassification means it predicted 100% as “English” as shown in the graph. The fifth word “Good” is recognized with 93% accuracy and misclassification is 07%, which means it predicted 93% as “Good” and 07% as “Bottle”, as in the graph. The sixth word “People” is recognized with 86% accuracy and misclassification is 14%, which means it predicted 86% as “People” and 14% as “Dog”, and “Today” as in the graph. The seventh word “Today” is recognized with 100% accuracy and no misclassification means it predicted 100% as “Today” as in the graph. Table 2 shows the classification report for an audio database. The precision, recall, accuracy, and F1-score of the proposed system are perceived as 91%, 90%, 90%, and 91% respectively.

5.2 Visual speech recognition evaluation

Figure 10 shows the number of epoch used for the visual model training model and here shows the training accuracy

```

Epoch 195/200
14/14[=====] - 0s 17ms/step - loss: 0.0941 - accuracy: 0.9651 - val_loss: 0.3379
Val_accuracy: 0.8952
Epoch 196/200
14/14[=====] - 0s 17ms/step - loss: 0.0992 - accuracy: 0.9634 - val_loss: 0.3087
Val_accuracy: 0.9143
Epoch 197/200
14/14[=====] - 0s 17ms/step - loss: 0.1130 - accuracy: 0.9649 - val_loss: 0.3203
Val_accuracy: 0.9143
Epoch 198/200
14/14[=====] - 0s 16ms/step - loss: 0.1252 - accuracy: 0.9596 - val_loss: 0.3632
Val_accuracy: 0.9143
Epoch 199/200
14/14[=====] - 0s 17ms/step - loss: 0.1355 - accuracy: 0.9487 - val_loss: 0.4554
Val_accuracy: 0.8952
Epoch 200/200
14/14[=====] - 0s 16ms/step - loss: 0.1254 - accuracy: 0.9662 - val_loss: 0.4187
Val_accuracy: 0.9048
    
```

Fig. 6 Training of Epochs for audio

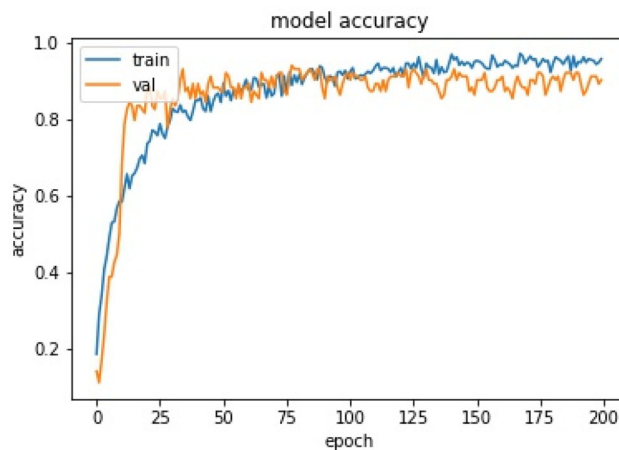


Fig. 7 Model accuracy curve for audio, epoch vs accuracy

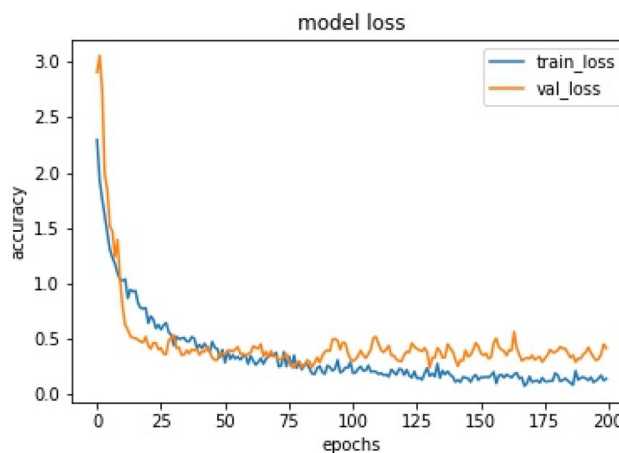


Fig. 8 Model loss curve for audio, epoch vs accuracy

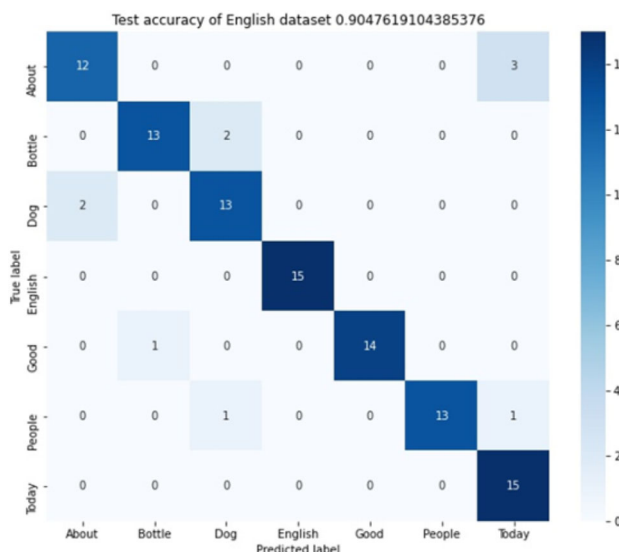


Fig. 9 Confusion matrix for audio-only model

Table 2 Classification report for audio model

	Precision	Recall	f1-score	Support
About	0.86	0.80	0.83	15
Bottle	0.93	0.87	0.90	15
Dog	0.81	0.87	0.84	15
English	1.00	1.00	1.00	15
Good	1.00	0.93	0.97	15
People	1.00	0.87	0.93	15
Today	0.79	1.00	0.88	15
Accuracy			0.90	105
Macro avg	0.91	0.90	0.91	105
Weighted avg	0.91	0.90	0.91	105

```

accuracy: 0.7143
Epoch 120/125
14/14 [=====] - 0s 25ms/step - loss: 0.4718 - accuracy: 0.8292 - val_loss: 1.0923 - val_
accuracy: 0.6667
Epoch 121/125
14/14 [=====] - 0s 23ms/step - loss: 0.6164 - accuracy: 0.7785 - val_loss: 1.0190 - val_
accuracy: 0.6762
Epoch 122/125
14/14 [=====] - 0s 25ms/step - loss: 0.4992 - accuracy: 0.8139 - val_loss: 1.0527 - val_
accuracy: 0.7048
Epoch 123/125
14/14 [=====] - 0s 23ms/step - loss: 0.4948 - accuracy: 0.8555 - val_loss: 1.1141 - val_
accuracy: 0.6381
Epoch 124/125
14/14 [=====] - 0s 23ms/step - loss: 0.4454 - accuracy: 0.8539 - val_loss: 1.0249 - val_
accuracy: 0.7143
Epoch 125/125
14/14 [=====] - 0s 23ms/step - loss: 0.4800 - accuracy: 0.8273 - val_loss: 0.9814 - val_
accuracy: 0.7143
    
```

Fig. 10 Training of Epochs for Visual Speech

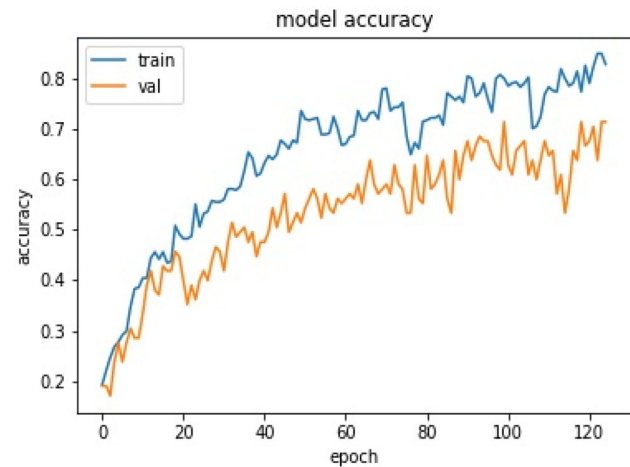


Fig. 11 Model accuracy curve for Visual Speech, epoch vs accuracy

71.43% and testing accuracy of 82.73%. Figure 11 shows the model accuracy model for the visual model and the graph is epoch versus accuracy. This graph shows the training and validation graph. Figure 12 shows the loss

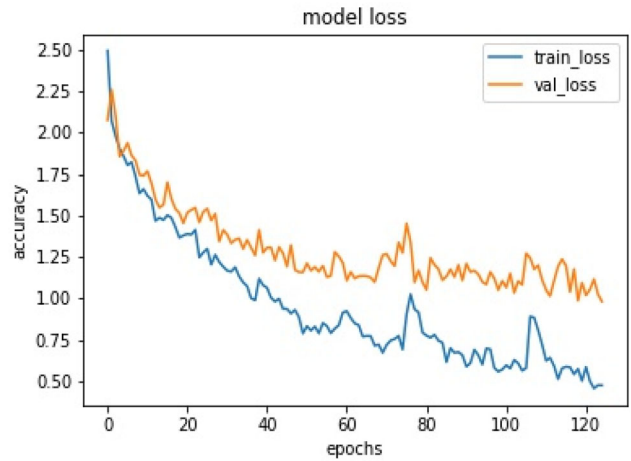


Fig. 12 Model loss curve for Visual Speech, epoch vs accuracy

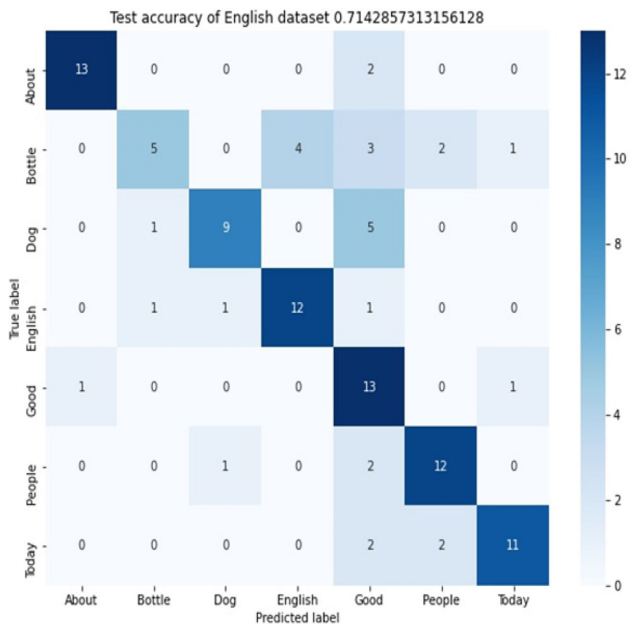


Fig. 13 Confusion matrix for Visual Speech Recognition

curve of the visual model and this graph shows the training loss and validation loss. Figure 13 shows the accuracy and misclassification of each word of visual speech recognition. The first word “About” is being recognized with 86% accuracy and misclassification is only 14%, which means the algorithm predicted 86% as “About” and 14% as “Good” as in the graph. The second word is “Bottle”. It is recognized with 33% accuracy and misclassification is 67%, which means the algorithm predicted 33% as “Bottle” and 26% as “English”, 20% as “Good”, 13% as “People”, and 6.6% as “Today” as in the graph. The third word “Dog” is recognized with 60% accuracy and

misclassification is 40%, which means it predicted 60% as “Dog” and 40% as “Bottle”, and “Good” as shown in the graph. The fourth word “English” is recognized with 80% accuracy and misclassification is 40%, which means it predicted 80% as “English” and 20% as “Dog”, and “Good” as shown in the graph. The fifth word “Good” is recognized with 86% accuracy and misclassification is 14%, which means it predicted 86% as “Good” and 14% as “About”, and “Today” as in the graph. The sixth word “People” is recognized with 80% accuracy and misclassification is 20%, which means it predicted 80% as “People” and 20% as “Dog”, and “Good” as in the graph. The seventh word “Today” is recognized with 73% accuracy and misclassification is 27%, which means it predicted 73% as “Today” and 27% as “Good”, and “People” as in the graph. Table 3 shows the classification report for an audio database. The precision, recall, accuracy, and F1-score of the proposed system are perceived as 75%, 71%, and 71% respectively.

5.3 Audiovisual speech recognition evaluation

Figure 14 shows the number of epoch used for the audiovisual model training model and here shows the training accuracy of 88.57% and testing accuracy of 91.93%. Figure 15 shows the model accuracy model for the audiovisual model and the graph is epoch versus accuracy. Figure 16 shows the loss curve of the audiovisual model and this graph shows the training loss and validation loss. Figure 17 shows the accuracy and misclassification of each word of visual speech recognition. The first word “About” is being recognized with 73% accuracy and misclassification is only 27%, which means the algorithm predicted 73% as “About” and 27% as “Bottle”, and “Dog” as in the graph.

Table 3 Classification report for visual model

	Precision	Recall	f1-score	Support
About	0.93	0.87	0.90	15
Bottle	0.71	0.33	0.45	15
Dog	0.82	0.60	0.69	15
English	0.75	0.80	0.77	15
Good	0.46	0.87	0.60	15
People	0.75	0.80	0.77	15
Today	0.85	0.73	0.79	15
accuracy			0.71	105
macro avg	0.75	0.71	0.71	105
Weighted avg	0.75	0.71	0.71	105

```

Epoch 225/230
14/14 [=====] - 1s 98ms/step - loss: 0.1980 - accuracy: 0.9342 - val_loss: 0.6441 - val_
accuracy: 0.8190
Epoch 226/230
14/14 [=====] - 1s 95ms/step - loss: 0.2188 - accuracy: 0.9309 - val_loss: 0.5878 - val_
accuracy: 0.8381
Epoch 227/230
14/14 [=====] - 1s 94ms/step - loss: 0.3590 - accuracy: 0.9272 - val_loss: 0.4332 - val_
accuracy: 0.8762
Epoch 228/230
14/14 [=====] - 1s 98ms/step - loss: 0.1792 - accuracy: 0.9432 - val_loss: 0.5169 - val_
accuracy: 0.8571
Epoch 229/230
14/14 [=====] - 1s 92ms/step - loss: 0.1760 - accuracy: 0.9428 - val_loss: 0.5601 - val_
accuracy: 0.8381
Epoch 230/230
14/14 [=====] - 1s 91ms/step - loss: 0.2748 - accuracy: 0.9193 - val_loss: 0.4987 - val_
accuracy: 0.8857
    
```

Fig. 14 Training of Epochs for Audiovisual Speech

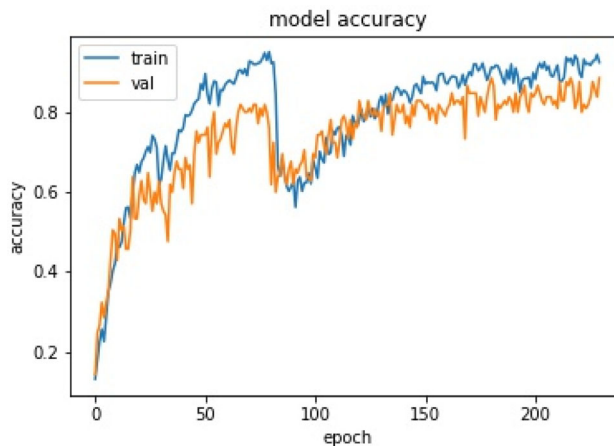


Fig. 15 Model accuracy curve for Audiovisual Speech, epoch vs accuracy

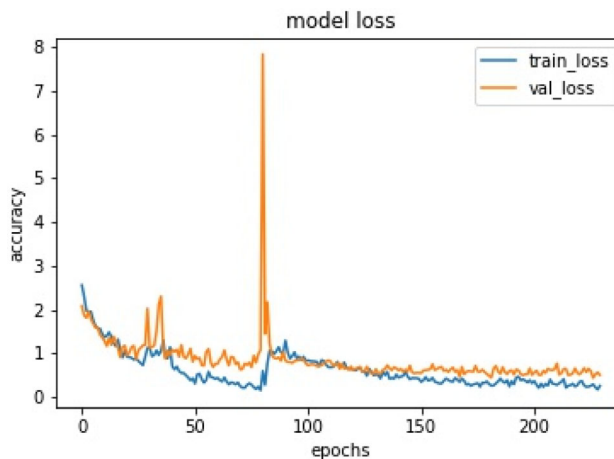


Fig. 16 Model loss curve for Audiovisual Speech, epoch vs accuracy

The second word is “Bottle”. It is recognized with 80% accuracy and misclassification is 20%, which means the algorithm predicted 80% as “Bottle” and 20% as “About”, “Dog”, and “English” as in the graph. The third word

Confusion Matrix, Normalized

True label	About	11.000	1.000	3.000	0.000	0.000	0.000	0.000
	Bottle	1.000	12.000	1.000	1.000	0.000	0.000	0.000
	Dog	0.000	0.000	15.000	0.000	0.000	0.000	0.000
	English	0.000	0.000	0.000	15.000	0.000	0.000	0.000
	Good	0.000	0.000	0.000	0.000	13.000	1.000	1.000
	People	0.000	0.000	0.000	0.000	0.000	13.000	2.000
	Today	0.000	0.000	0.000	1.000	0.120	0.000	14.000
		Predicted label	About	Bottle	Dog	English	Good	People

accuracy =0.91, misclass=0.0990

Fig. 17 Confusion matrix for Audiovisual Speech Recognition

Table 4 Classification report for audiovisual model

	Precision	Recall	f1-score	Support
About	0.92	0.73	0.81	15
Bottle	0.92	0.80	0.86	15
Dog	0.79	1.00	0.88	15
English	0.88	1.00	0.94	15
Good	1.00	0.87	0.93	15
People	0.93	0.87	0.90	15
Today	0.82	0.93	0.87	15
accuracy			0.91	105
macro avg	0.91	0.91	0.91	105
Weighted avg	0.91	0.91	0.91	105

“Dog” is recognized with 100% accuracy and no misclassification means it predicted 100% as “Dog” as shown in the graph. The fourth word “English” is recognized with 100% accuracy and no misclassification means it predicted

100% as “English” as shown in the graph. The fifth word “Good” is recognized with 86% accuracy and misclassification is 14%, which means it predicted 86% as “Good” and 14% as “People”, and “Today” as in the graph. The sixth word “People” is recognized with 86% accuracy and misclassification is 14%, which means it predicted 86% as “People” and 14% as “Today”, as in the graph. The seventh word “Today” is recognized with 93% accuracy and misclassification is 07%, which means it predicted 93% as “Today” and 07% as “English” as in the graph. Table 4 shows the classification report for an audio database. The precision, recall, accuracy, and F1-score of the proposed system are perceived as 89%, 89%, 89%, and 88% respectively. Table 5 shows the comparison of the existing output with proposed methods with accuracy. Table 6 shows the comparison of the existing output of audiovisual speech recognition with proposed methods with accuracy.

6 Conclusion

In this work, we develop audiovisual speech recognition for a custom dataset and the dataset contains English words. First, we extract the audio features from the video and use 1D CNN for classification and got 90% accuracy and recognition of visual speech using the LSTM technique and got 71.42% accuracy. When combined the audio and visual using a deep neural network to get better accuracy in the AVSR model. The combined audio and video involving deep neural networks got 91% accuracy. Limitations are the proposed AVSR model recognizes a single word, this model cannot recognize sentences, and this does not end to end model. In future work, we can use more datasets for training and testing and plan to use different neural networks. Create a database in different angles other than the straight to the face to the speaker.

Table 5 Obtained visual results on the custom dataset in comparison with the existing method

Method	HMM [25]	DCNN [26]	CNN &LSTM RNN [28]	LSTM
Accuracy	46.6%	59.23%	57.7%	71.42%
Dataset	AV Letters	AV Letters	AV Letters	CUSTOM

Table 6 Obtained audiovisual results on a custom dataset in comparison with the existing method

Method	MFFCN [28]	CNN & LSTM RNN [29]	DCNN [Proposed]
Accuracy	82.7%	84.4%	91%
Dataset	TCD-TIMIT	AV Letters	Custom

Funding Not applicable.

Availability of data and material The data that support the findings of this study are available from the corresponding author on reasonable request.

Code availability The code available on reasonable request to the corresponding author.

Declarations

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

References

- Afouros T, Chung JS, Senior A, Vinyals O, Zisserman A (2018) Deep audio-visual speech recognition. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2018.2889052>
- Ivo I (2011) Speech and language technologies, pp 285–289, <https://doi.org/10.5772/938>
- Shaikh AA, Kumar DK (2011) Visual speech recognition using optical flow and support vector machines. *Int J Comput Intell Appl* 10:171. <https://doi.org/10.1142/S1469026811003045>
- Stafylakis T, Tzimiropoulos G (2017) Combining residual networks with LSTMs for lip reading <https://doi.org/10.21437/InterSpeech.2017-85>
- Shillingford B, Assael YM, Hoffman MW, Paine TL, Hughes C, Prabhu U, Liao H, Sak H, Rao K, Bennett L, Mulville M, Coppin B, Laurie B, Senior AW, Freitas ND (2019) Large-scale visual speech recognition. arXiv: abs/1807.05162
- Courtney L, Sreenivas R (2019) Learning from videos with deep convolutional LSTM networks. arXiv preprint. arXiv: 1904.04817.
- Sterpu G, Saam C, Harte N (2018) Can DNNs learn to lipread full sentences? In: 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp 16–20, <https://doi.org/10.1109/ICIP.2018.8451388>
- Kumar Y, Jain R, Salik K, Shah RR, Yin Y, Zimmermann R (2019) Lipper: synthesizing thy speech using multi-view lipreading. In: Proceedings of the AAAI Conference on artificial intelligence. 33: 2588–2595, <https://doi.org/10.1609/aaai.v33i01.33012588>
- Xu K, Li D, Cassimatis N, Wang X (2018) LCArNet: End-to-end lipreading with cascaded attention-CTC. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, 2018, pp 548–5558, <https://doi.org/10.1109/FG.2018.00088>
- Margam D, Aralikatti R, Sharma T, Thanda A, Pujitha K, Roy S, Venkatesan S (2019) LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models. Published in ArXiv 2019. <https://dblp.org/db/journals/corr/corr1906.html#abs-1906-12170>. Accessed Jan 2021
- Stafylakis T, Khan MH, Tzimiropoulos G (2018) Pushing the boundaries of audiovisual word recognition using Residual Networks and LSTMs. *Comput Vis Image Underst* Vol 176–177:22–32. <https://doi.org/10.1016/j.cviu.2018.10.003>
- Zhang S, Lei M, Ma B, Xie L (2019) Robust audio-visual speech recognition using bimodal Dfsmn with multi-condition training and dropout regularization. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp 6570–6574, <https://doi.org/10.1109/ICASSP.2019.8682566>
- Noda K, Yamaguchi Y, Nakadai K, Okuno H, Ogata T (2014) Audio-visual speech recognition using deep learning. *Appl Intell* 42:722–737. <https://doi.org/10.1007/s10489-014-0629-7>
- Petridis S, Li Z, Pantic M (2017) End-to-end visual speech recognition with LSTMS. In: IEEE International Conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, 2017, pp 2592–2596, <https://doi.org/10.1109/ICASSP.2017.7952625>
- Tao F, Busso C (2018) Gating neural network for large vocabulary audiovisual speech recognition. *IEEE/ACM Trans Audio Speech Lang Process* 26(7):1290–1302. <https://doi.org/10.1109/TASLP.2018.2815268>
- Petridis S, Stafylakis T, Ma P, Tzimiropoulos G, Pantic M (2018) Audio-visual speech recognition with a hybrid CTC/attention architecture. In: IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp 513–520, <https://doi.org/10.1109/SLT.2018.8639643>
- Goh Y, Lau K, Lee Y (2019) Audio-visual speech recognition system using recurrent neural network. In: 2019 4th International Conference on information technology (InCIT), Bangkok, Thailand, 2019, pp. 38–43, <https://doi.org/10.1109/INCIT.2019.8912049>
- Wang J, Wang L, Zhang J, Wei J, Yu M, Yu R (2018) A large-scale depth-based multimodal audio-visual corpus in mandarin. In: IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, United Kingdom, 2018, pp. 881–885, <https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00146>
- Ochiai T, Delcroix M, Kinoshita K, Ogawa A, Nakatani T (2019) Multimodal SpeakerBeam: single channel target speech extraction with audio-visual speaker clues. In: Interspeech 2019, pp 2718–2722, <https://doi.org/10.21437/interspeech.2019-1513>
- Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. In: IEEE Conference on computer vision and pattern recognition (CVPR), Honolulu, HI, 2017, pp 3444–3453, <https://doi.org/10.1109/CVPR.2017.367>
- Jha V, Namboodiri P, Jawahar CV (2018) Word spotting in silent lip videos. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 2018, pp 150–159, <https://doi.org/10.1109/WACV.2018.00023>
- Thabet Z, Nabih A, Azmi K, Samy Y, Khoriba G, Elshehaly M (2018) Lipreading using a comparative machine learning approach. In: (2018) First International Workshop on Deep and Representation Learning (IWDRL), Cairo, 2018, pp 19–25, <https://doi.org/10.1109/IWDRL.2018.8358210>
- Kumar Y, Aggarwal M, Nawal P, Satoh S, Ratn Shah R, Zimmermann R (2018) Harnessing AI for speech reconstruction using multi-view silent video feed. In: 2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, p 9. <https://doi.org/10.1145/3240508.3241911>.
- Lu Y, Liu Q (2018) Lip segmentation using automatic selected initial contours based on localized active contour model. *Eurasip J Image Video Process* 1:2018. <https://doi.org/10.1186/s13640-017-0243-9>
- Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R (2002) Extraction of visual features for lipreading. *IEEE Trans Pattern Anal Mach Intell* 24:198–213. <https://doi.org/10.1109/34.982900>
- Mesbah A, Hammouchi H, Berrahou A et al (2019) Lip Reading with Hahn convolutional neural networks moments. *Image Vis Comput* 88:76–83. <https://doi.org/10.1016/j.imavis.2019.04.010>

27. Shashidhar R, Patilkulkarni S (2021) Visual speech recognition for small scale dataset using VGG16 convolution neural network. *Multimed Tools Appl* 15:14. <https://doi.org/10.1007/s11042-021-11119-0>
28. Xu X, Xu D, Jia J, Wang Y, Chen B (2021) MFFCN: multi-layer feature fusion convolution network for audio-visual speech enhancement. *arXiv: abs/2101.05975*
29. Feng W, Guan N, Li Y, Zhang X, Luo Z (2017) Audio visual speech recognition with multimodal recurrent neural networks. In: 2017 International Joint Conference on neural networks (IJCNN), 2017, pp. 681–688, <https://doi.org/10.1109/IJCNN.2017.7965918>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.