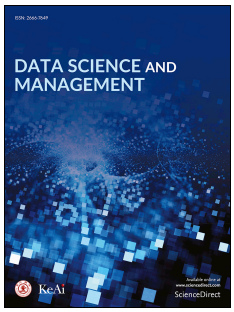
Journal Pre-proof



Audiovisual speech recognition based on a deep convolutional neural network

R. Shashidhar, S. Patilkulkarni, Vinayakumar Ravi, H.L. Gururaj, Moez Krichen

PII: S2666-7649(23)00045-0

DOI: https://doi.org/10.1016/j.dsm.2023.10.002

Reference: DSM 74

To appear in: *Data Science and Management*

Received Date: 3 May 2023

Revised Date: 1 October 2023

Accepted Date: 5 October 2023

Please cite this article as: Shashidhar, R., Patilkulkarni, S., Ravi, V., Gururaj, H.L., Krichen, M., Audiovisual speech recognition based on a deep convolutional neural network, *Data Science and Management* (2023), doi: https://doi.org/10.1016/j.dsm.2023.10.002.

Deep convolutional neural network based Audio Visual Speech Recognition

[1]Shashidhar R , [2]S Patilkulkarni, [3*]Vinayakumar Ravi, [4*]Gururaj HL, [5 6]Moez Krichen

[1,2] Department of Electronics and Communication Engineering, JSS Science and Technology University, Mysuru, India-570006

[3]Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar 34754, Saudi Arabia vravi@pmu.edu.sa

[4]Department of Information Technology, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India gururaj.hl@manipal.edu

[5]Department of Information Technology, Faculty of Computer Science and Information Technology (FCSIT), Al-Baha University, Alaqiq 65779-7738, Saudi Arabia

[6]ReDCAD Laboratory, University of Sfax, Sfax 3038, Tunisia

Corresponding Author: Vinayakumar Ravi (vravi@pmu.edu.sa) and Gururaj HL (gururaj.hl@manipal.edu)

Audiovisual speech recognition based on a deep convolutional neural network

Abstract: Audiovisual speech recognition is an emerging research topic. Lipreading is the recognition of what someone is saying using visual information, primarily lip movements. In this study, we created a custom dataset for Indian English linguistics and categorized it into three main categories: (1) audio recognition, (2) visual feature extraction, and (3) combined audio and visual recognition. Audio features were extracted using the mel-frequency cepstral coefficient, and classification was performed using a one-dimension convolutional neural network. Visual feature extraction uses Dlib and then classifies visual speech using a long short-term memory type of recurrent neural networks. Finally, integration was performed using a deep convolutional network. The audio speech of Indian English was successfully recognized with accuracies of 93.67% and 91.53%, respectively, using testing data from two hundred epochs. The training accuracy for visual speech recognition using the Indian English dataset was 77.48% and the test accuracy was 76.19% using 60 epochs. After integration, the accuracies of audiovisual speech recognition using the Indian English dataset for training and testing were 94.67% and 91.75%, respectively.

Keywords: Audiovisual speech recognition, Custom dataset, 1D CNN, DCNN, LSTM, Lipreading, Dlib, MFCC

1. Introduction

With recent technological advancements, pattern recognition has become vital, as it enables computers to imitate how the human brain perceives the environment (Abderrahim et al., 2019). Compared to other recognition systems, such as fingerprint, gesture, and facial identification, audiovisual speech recognition is more advantageous and reliable, making it a fundamental part of the human–computer interface. Lipreading(Abdelwahab and Busso, 2019], which involves analyzing video data to interpret the movements of the lips and other facial features, requires several essential processes such as pattern recognition, image processing, and computer vision. These processes enable developers and researchers to create more precise and dependable audiovisual speech recognition systems for various applications, including assistive technologies and security systems. Audiovisual speech recognition has the advantage of lipreading and image processing capabilities that help speech recognition algorithms to identify ambivalent words more reliably. Therefore, building an audiovisual speech recognition model to translate audio and visual inputs is the main goal. To aid future studies in this area, this study evaluated the performance and restrictions of a hybrid model employed for audiovisual speech recognition and the results are detailed in this paper.

The motivation for the proposed work is to help people with impaired hearing. People with hearing impairment often rely on lipreading to understand others in noisy environments or when audio signals are unclear. However, lipreading requires training and it can be challenging for hearing-impaired individuals to detect spoken words. Audiovisual speech recognition, which includes the interpretation of both audio and visual signals, can be the solution for this problem and help people with hearing impairments participate in conversations. The goal of this project is to develop an algorithm for audiovisual speech recognition that predicts spoken words from videos without audio, and vice versa, to assist people with hearing impairment.

The following are the primary goals of this research work:

- Create an Indian English language database.
- Develop a visual speech recognition algorithm.
- Improve feature extraction and pattern learning by altering the neural network.
- Obtain the finest audiovisual speech recognition results possible.

The scientific significance of the anticipated models encompasses multimodel understanding, neuroscientific insights, and cross-model learning, whereas the practical significance of the anticipated models includes improved speech recognition accuracy, accessibility, security, and authentication.

The remainder of this paper is organized as follows: Section 2 discusses the existing audio visual speech recognition methodologies and tools. Section 3 describes the proposed technique, which is broken down into three basic models: the auditory, visual, and fusion models. The findings are described in detail in Section 4. Projected and subsequent works are described in Section 5.

2. Audio visual speech recognition tools

This section discusses previous works related to audiovisual and visual speech recognition, including a review of the different datasets used in these studies and the feature extraction and classification strategies employed. Additionally, this section provides insights into the current state of knowledge in this field and identifies areas for future research. A hybrid algorithm for audiovisual voice recognition was proposed (Debnath and Roy, 2021). The VISWa database was subjected to multiclass support vector machine and naïve Bayes classification. An innovative Hahn Convolutional Neural Network design was proposed (Abderrahim et al., 2019) and they used it to analyze the AVLetters, OuluVS2, and BBCLRW datasets. The CRSS-4ENGLISH-14 corpus dataset was used by Tao et al. They used a deep neural network with a hidden Marko model (HMM) as well as a Gaussian mixture model with a hidden Marko model to achieve greater accuracy(Abdelwahab and Busso., 2019). (Wu et al., 2019) introduced a time-domain speech separation network for the Lip Reading Sentences 2 (LRS2) dataset. They used an audio encoder for audio processing, a video encoder for lipreading, and an audiovisual speech recognition system for audiovisual speech recognition. Their approach involved using these different components to analyze and separate speech signals from a video, thereby enabling more accurate speech recognition in noisy environments.

Using the Lip Reading in the Wild (LRW) dataset, (Petridis et al. 2018) presented audiovisual speech recognition as an end-to-end process. They used residual neural networks and bi-directionally gated recurrent units. (Goh et al.,2019) used a custom dataset for their work, with 50 males and 50 females as representatives, and employed recurrent neural networks (RNNs) to recognize audiovisual speech from 5,000 visual speech samples and 1,000 audio samples with a two-second duration. (Sooraj et al. 2020)studied different methods of lipreading, mainly focusing on speech processing. Lipreading techniques use deep learning and machine learning. The main aspects of visual speech are lip detection, lip feature extraction, classification methods, challenges, and different open-source database details. (Belete,2019) used the Viola-Jones algorithm for lip detection and mouth ROI extraction and DWT for feature extraction. The HMM was used as a classifier for visual speech recognition, mel-frequency cepstral coefficient (MFCC) was used, and the HMM was used as an audio classifier. The integrated audiovisual speech, through coupled hidden Markov model (CHMM), was used as the classifier. (Afouras et al.2018) proposed sentence recognition and compared two lipreading methods. The authors worked on the audio when it was noisy as well and they introduced new and publicly available datasets, like the LRS2, which holds thousands of natural audio and spoken sentences from BBC television programs. (Afouras et al. 2018)introduced a novel multimodal database called the LRS3-TED, which includes 400 hours of TEDx videos and is one of the benchmark datasets used for visual speech recognition.

The LRS3-TED dataset, for which (Makino et al.,2019) used a RNN transducer, is available on YouTube. It is TEDx videos and compared with other large datasets called YTDEV18. (Tao and Busso,2018) used the CRSS-4ENGLISH-14 database, which contains 61 hours of speech from 155 people. The authors trained the dataset using deep neural networks and HMMs. The LRS2 database was used by (Yu et al.,2021) for audiovisual speech recognition. To integrate the auditory and visual data, they applied a hybrid approach. (Tan et al. 2020) used a multistage multimodal network to perform audiovisual voice detection. Their experimental findings demonstrated improved output and accuracy when the two models were trained independently.

(Martinez et al. 2020) proposed a bidirectional gated recurrent unit that used the LRW and LRW1000 datasets; the proposed model exhibited improvements of 1.2% and 3.2%, respectively. (Meutzner et al. 2017)proposed a state-based approach for integrating audio and visual speech in audiovisual speech recognition. They suggested using a deep neural network to combine audio and visual information to improve recognition accuracy. This technique involves modeling audio and visual streams separately and then integrating them using a state-based approach. Meutzner et al.'s method is one of the most recent techniques in this field and has shown promising results in improving audiovisual speech recognition accuracy. (Yuan et al. 2018) recommended a multimodal gated recurrent units for the audiovisual recognition of speech systems. It includes three main steps: the initial stage extracts features from the database, the second step is data augmentation, and the third step is integration and recognition. (Tao and Busso,2021) used a sizable audiovisual archive with a duration of approximately 60 hours as well as a database with various surroundings and data gathered from various channels. (Feng et al. 2017) developed a multimodal RNN for audiovisual speech recognition. Their system consisted of three primary parts: the audio component, the visual component, and the integration of the audio and visual parts through a RNN. The

audio and visual components were modeled separately and then fused at the neural network level to improve the recognition accuracy. This approach was shown to be effective in recognizing speech in noisy and challenging environments, making it a promising technique for various applications such as human–robot interactions and speech-enabled devices. (Agarwal et al. 2021)used two publicly accessible datasets, GRID and TCD-TIMIT, as well as specialized Asian databases. From these datasets, the authors selected nine hundred random videos.

(Stafylakis and Tzimiropoulos,2017) used a BBCLRW dataset and a residual network with long short-term memory (LSTM) for training and testing, from which they achieved better accuracy. (Li et al.2019)worked on lipreading using a deep neural network with multiple models, dynamic vision, and dynamic audio sensors. First, they trained each model individually, then they combined the models and trained the network. (Nadeemhashmi et al.2018) proposed a 12-layer convolution neural network (CNN) for lipreading on a MIRACLE-VC1 dataset. (Santos and Abel, 2019) worked on various deep neural networks for visual speech recognition to obtain better results and outcomes. (Jang et al.2019) worked on lipreading in two different ways: first, they used complete lip images and then they used spots around lip benchmarks. They used the OuluVS2 dataset and applied the VGG-m neural network to improve accuracy. (Tao and Busso,2018) proposed an LSTM for audiovisual activity detection and trained a bimodal RNN with an LSTM layer.

(Wand and Vu,2018) used the GRID dataset to train and test a deep neural network. (Wei et al. 2018)suggested a novel system for lipreading called a densely associated convolution network, which captures visual representations from color images. The authors worked on 3D lip physiological features based on the position and structure of the face. (Petridis et al. 2018) proposed a hybrid algorithm for audiovisual recognition. The authors applied CTC and the attention architecture to the LRS2 database for audiovisual speech recognition.

(Jadczyk,2018)conducted a Polish audiovisual voice recognition project wherein the work was divided into three subcategories: audio, visual speech, and integration. For the audio, they employed MFCC to extract the features, an HMM to extract the lip features, and multistream HMM for incorporation. (Shashidhar and Patilkulkarni,2021) worked on a traditional lipreading database. The authors used a custom 250-item dataset and applied a pre-trained model called VGG16 for better accuracy. (Cornejo and Pedrini,2019) proposed a deep CNN (DCNN) for audiovisual voice detection. Their approach first involved separating the audio from the video and then extracting audio data using two-dimensional CNN. They also extracted visual speech using principal component analysis and linear discriminant analysis and translated it into the census transform. Finally, audio and visual feature extractions were merged to improve the accuracy of audiovisual voice detection. Their method showed promising results on various benchmark datasets, demonstrating the effectiveness of DCNNs for audiovisual voice detection(Shashidhar and Sudarshan, 2022).

Additionally, authors (Shashidhar et al., 2022)created a specific dataset for the Kannada linguistics and used a feed-forward CNN to integrate audio visual and speech recognition.

The authors introduced a new approach called a multimodal sparse transformer network (MMST). This method employs a sparse self-attention mechanism to focus selectively on important parts of the data and thus improve global information processing (Dupont and Luettin,2000) demonstrated the high performance of their proposed system on a large database of continuously spoken digits involving multiple speakers. The authors used a combination of the MSHMM, denoised MFCCs, and visual features to improve the results of multimodal isolated word recognition. They demonstrated that this approach can lead to better word recognition accuracy than using only audio features(Nefian et al. 2002).

(Noda et al. 2015)presented two statistical models, namely the CHMM and factorial HMM (FHMM), for audiovisual integration. The performance of AV speech recognition varies greatly, both in terms of the overall recognition score and the amount of audiovisual gain(Ken et al.1998) To address this variability, consonant confusion was analyzed based on phonetic features to determine the level of redundancy(Heckmann et al.2002) The authors conducted a recognition task using manually controlled noise to evaluate the performances of various weighting schemes in neural networks, all of which were trained using clean data(Song et al.2022) The inclusion of the visual modality in AV Taris outperformed the audio-only version of Taris, indicating the effectiveness of utilizing visual information in speech recognition within Taris's real-time decoding framework (George and Naomi, 2022).

The main limitations or disadvantages of the existing methods are the use of existing data, such as YouTube and TEDx video datasets, which pose dependence on visual cues, complexities, and limited datasets. To overcome these limitations, we developed custom datasets for a natural environment with a signal-to-noise ratio of 20db.

3. Proposed method

In this section, we describe how we created a dataset, extracted and classified features from an audio model, extracted and classified features from a visual model, and integrated the visual and audio components using a DCNN. Fig. 1 shows the proposed block diagram of this process.

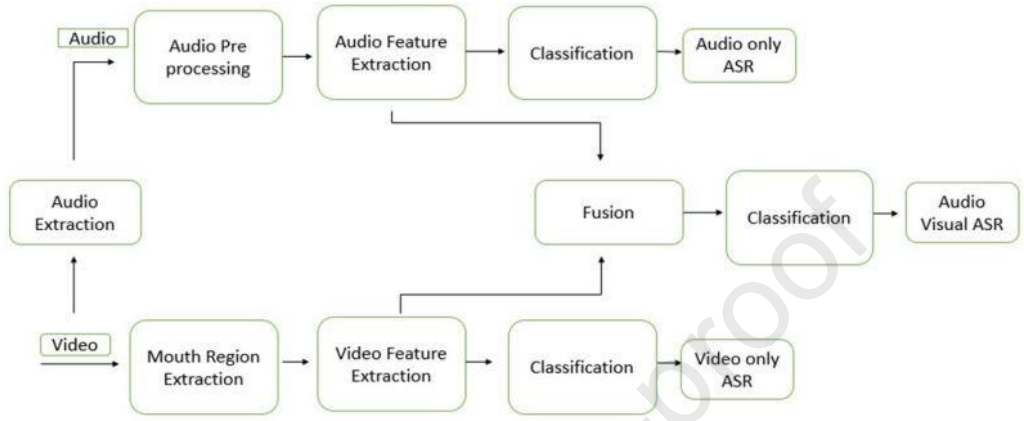


Fig. 1. Proposed block diagram

3.1 Dataset Formation

The software was trained to recognize lip-gesture sequences using a movie database in which users list specific words. A 4K professional-grade video recorder, movable lights, and a controlled, isolated environment were used to record the dataset. The authors used a high-definition camera with a resolution of 1080 × 1920 pixels to create their dataset. They were able to capture an average of 80–100 frames per video with a duration of 1.6 seconds recorded at a rate of 60 fps. This amount of data was sufficient for conducting the processes used in our research. The typical video size for one person was 10 MB.

The dataset used in the study consisted of 80 videos captured from 16 participants (eight males and eight females aged between 18 and 30 years). Recordings were collected in a controlled and quiet environment with minimal background noise. The purpose of collecting these data was to develop and evaluate the lipreading process using tools such as speech recognition. Each of the 16 participants spoke eight words, which were recorded five times. Consequently, the dataset consisted of 640 bytes of information (eight words × five times × 16 persons).

Here, nine English words—“About,” “Bad,” “Bottle,” “Come,” “Cow,” “Good,” “Pencil,” “Read,” and “Where”—were used. These words were selected randomly from an average of eighty-odd movies. The videos in the dataset were edited using a Microsoft video editor to include only one second of speech for each word. The video frame rate was adjusted to 30 fps. Two sets of datasets were created: a training set comprising 75% of the total data and a validation set comprising 25% or 21 videos. The training set was used to train the model and the validation set was used for model evaluation.

In formation of the database videos are recorded at a resolution of 1080 pixles with a smooth frame rate of 60 Frams Per Second, resulting in high-quality footage. Each video has an average duration of 1 to 1.20 seconds, and their file size typically falls around 10 megabytes.

3.2 Acoustic Model

The video data were first used to generate audio files, which were then saved in .wav format using the FFmpeg tool. The open-source Python library Librosa was then used to extract audio features, using MFCC to extract features from the audio that can be more easily categorized. These features were combined to create a 193 × 1 feature vector. This vector was then passed through two dense layers, followed by a CNN consisting of one Conv1D layer, one MaxPooling 1D layer, one batch normalization layer, and one dropout layer. One-dimensional CNNs (1D CNNs) are effective in processing signals with sequences in one dimension, making them useful for audio signal analysis. The outputs of the associated layers of the audio model were then obtained as follow:

$$a^1 - b^1 + \sum_{i=1}^{193}(Conv1D(W_i, X\_train\_audio[i]) \quad (1)$$

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

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

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

$$y^4 = R(w^2 y^2 + b^3) \quad (5)$$

In the proposed model, the “layer i” vector is represented by yi and the ReLu activation function is denoted by R. The weights and biases of layer i are denoted by Wi and bi, respectively. A softmax layer was added to the network for classification. The model was trained using cross-entropy, which was expressed as a loss function.

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

3.3 Visual model

To extract the initial mouth region from the video, the researchers used the Dlib library available in Python 3, as depicted in Fig. 2. The regions were then converted to grayscale to reduce the complexity of the model, as shown in Fig. 3. The outer lip location was then obtained and stored in the feature vector. Next, a deep LSTM network was created, which is a model that includes a network of LSTMs and dense layers. An LSTM type of RNN (LSTM RNN) can learn order dependence in sequence prediction problems. The three gates of the LSTM are forget, input, and output gates.

Fig. 2. Extraction of mouth ROI

Fig. 3. Grayscale conversion

The first layer of the deep LSTM network model contained an LSTM layer with 128 hidden units and 8 time stamps. The internal workings of the LSTM cell can be broken down into three gates: the input gate, which measures the importance of the incoming data; the forget gate, which decides whether to retain or forget information from previous time steps; and the output gate, which determines the most relevant output to generate.

$$\hat{c}^t = tanh(W_c[a^t, x^t] + b_c) \quad (7)$$

$$\Gamma_u = \sigma(W_u[a^{t-1}, x^t] + b_u) \quad (8)$$

$$\Gamma_f = \sigma(W_f[a^{t-1}, x^t] + b_f) \quad (9)$$

$$\Gamma_o = \sigma(W_o[a^{t-1}, x^t] + b_o) \quad (10)$$

$$c^t = \Gamma_u * \hat{c}^t + \Gamma_f * c^{t-1} \quad (11)$$

$$a^t = \Gamma_o \tanh(c^t) \quad (12)$$

where c represents the memory cell, t represents the time stamp, and Γ_u represents the update gate. The forget gate is represented by Γ_f, and the output gate by Γ_o. W_f stands for the forget gate weights, W_o for the output gate weights, b for bias, and ct for contestant cell variables.

Subsequently, another LSTM layer was added, which resulted in the output of the second layer being as follow:

$$\widehat{c1}^t = tanh(W1_c[a1^t, x^t] + b1_c) \quad (13)$$

$$\Gamma 1_u = \sigma(W1_u[a1^{t-1}, x1^t] + b1_u) \quad (14)$$

$$\Gamma 1_f = \sigma(W1_f[a1^{<t-1>}, x1^{<t>}] + b1_f) \quad (15)$$

$$\Gamma 1_o = \sigma(W1_o[a1^{<t-1>}, x1^{<t>}] + b1_o) \quad (16)$$

$$c1^{<t>} = \Gamma 1_u * \widehat{c1}^{<t>} + \Gamma 1_f * c1^{<t-1>} \quad (17)$$

$$a1^{<t>} = \Gamma 1_o \tanh(c1^{<t>}) \quad (18)$$

The next three layers were very thick. These three dense layers produced the following output equations:

$$y^2 = R(W2 * a1 + a2) \quad (19)$$

$$y^3 = R(W3 * y^2 + b3) \quad (20)$$

$$y^4 = R(W4 * y^3 + b4) \quad (21)$$

Finally, this network had a softmax layer added for categorization. The cross-entropy, a function given by, serves as the loss function used to train the model.

$$L^{<t>}(\hat{y}^{<t>}) = y^{<t>} \log(\hat{y}^{<t>}) - (1 - y^{<t>}) log(1 - \hat{y}^{<t>}) \quad (22)$$

3.4 Combination of acoustic and visual model

The fusion model consists of three components: the auditory, visual, and audiovisual speech recognition models. The audio-only portion uses the same feature extraction method as the audio model, followed by building a DCNN with a softmax layer. The equations for this can be expressed in a manner similar to that of an audio model as follow:

$$a^1 = b^1 + \sum_{i=1}^{193} (Conv1D(W_{i,} X\_train\_audio[i]) \quad (23)$$

$$y^1 = R(a^1) \quad (24)$$

$$y^2 = R(w^2 y^1 + b^2) \quad (25)$$

$$y^3 = R(w^3 y^2 + b^3) \quad (26)$$

$$y^4 = R(w^4 y^3 + b^4) \quad (27)$$

$$y^5 = R(w^5 y^4 + b^5) \quad (28)$$

Similar to the video model, the video features were retrieved in the visual-only model. Subsequently, a deep LSTM network was built. A total of 128 unseen units with 8-time imprints were present in the top LSTM layer.

$$\hat{c}^{<t>} = tanh(W_c[a^{<t>}, x^{<t>}] + b_c) \quad (29)$$

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u) \quad (30)$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f) \quad (31)$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o) \quad (32)$$

$$c^{<t>} = \Gamma_u * \hat{c}^{<t>} + \Gamma_f * c^{<t-1>} \quad (33)$$

$$a^{<t>} = \Gamma_o \tanh(c^{<t>}) \quad (34)$$

These were monitored through a batch normalization layer, a dropout layer, and another LSTM layer with eight timestamps and 128 hidden units.

$$\widehat{c1}^{<t>} = tanh(W1_c[a1^{<t>}, x^{<t>}] + b1_c) \quad (35)$$

$$\Gamma 1_u = \sigma(W1_u[a1^{<t-1>}, x1^{<t>}] + b1_u) \quad (36)$$

$$\Gamma 1_f = \sigma(W1_f[a1^{<t-1>}, x1^{<t>}] + b1_f) \quad (37)$$

$$\Gamma 1_o = \sigma(W1_o[a1^{<t-1>}, x1^{<t>}] + b1_o) \quad (38)$$

$$c1^{<t>} = \Gamma 1_u * \widehat{c1}^{<t>} + \Gamma 1_f * c1^{<t-1>} \quad (39)$$

$$a1^{<t>} = \Gamma 1_o \tanh(c1^{<t>}) \quad (40)$$

Dropout and dense layers occur after the LSTM layer. Hence, the resultant equation for the dense layer is as follows:

$$y_v = R(W2 * a1^{<t>} + b2) \quad (41)$$

The feature map from the first dense layer of the audio-only portion and the feature map from the first LSTM layer of the visual-only portion were integrated. Subsequently, using Eqs. (3) and (12), we obtained

$$a_c = [a1^{<t>}, y^2] \quad (42)$$

The three thick layers of the DCNN were fed with the resulting feature map as input. A batch normalization layer and dropout layer were generated after the first two dense layers. The generated feature map was fed forward and provided as input to the DCNN with three dense layers. A batch normalization layer and dropout layer were generated after the first two dense layers. A fundamental feedforward neural network without loops is a DCNN. Fig. 4 depicts the architecture of a DCNN.

CNNs are a class of artificial neural networks inspired by the human visual system. They have proven to be extraordinarily effective in solving complex computer vision tasks, making them the cornerstone of modern machine learning and artificial intelligence. DCNNs are composed of multiple layers, each with a specific role in extracting and transforming visual features from input data.

One of the strengths of DCNNs is their ability to automatically learn the hierarchical representations of features. The lower layers capture simple features such as edges and corners, whereas the higher layers combine these features to recognize more complex patterns, ultimately leading to object recognition. DCNNs are trained using large datasets, which is typically supervised learning. This process involves feeding labeled images into the network and adjusting their internal parameters (weights and biases) through backpropagation and gradient descent. The goal is to minimize the loss function, making the network's predictions as close as possible to the true labels. DCNNs have demonstrated the ability to transfer knowledge learned from one task or dataset to another. This is particularly useful when working with limited labeled data, as pretrained models can be fine-tuned for specific tasks.

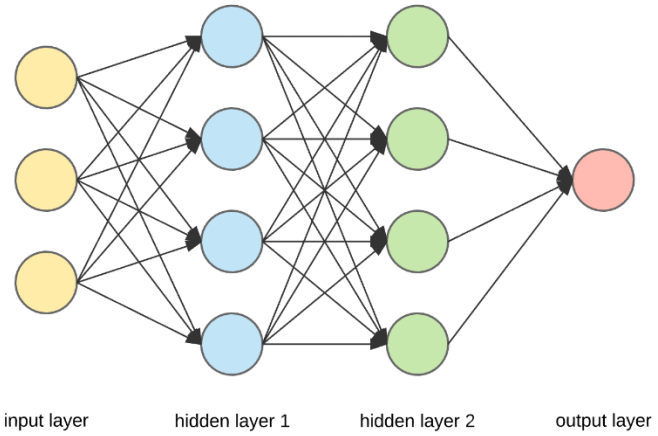


Fig. 4. The structure of deep convolutional neural networks

Equations for the output layer are:

$$y_{d1} = R(w_{d1} * a_c + b_{d1}) \tag{43}$$

$$y_{d2} = R(w_{d2} * y_{d1} + b_{d2}) \tag{44}$$

$$y_{d3} = R(w_{d3} * y_{d2} + b_{d3}) \tag{45}$$

Finally, all three steps are combined, resulting in a vector that integrates the output vectors of the aforementioned three steps. Hence, based on Eqs. (28), (41), and (45), we obtained the following:

$$a_{c2} = [y^5, y_v, y_{d3}] \tag{46}$$

The feature vector obtained from the audio-only model was fed as input to the DCNN. This network consists of three dense layers, followed by a batch normalization layer and a dropout layer.

$$y_{c1} = R(W_{c1} * a_{c2} + b_{c1}) \tag{47}$$

$$y_{c2} = R(W_{c2} * y_{c1} + b_{c2}) \tag{48}$$

$$y_{c3} = R(W_{c3} * y_{c2} + b_{c3}) \tag{49}$$

4. Result and discussion

This section discusses the evaluation and comparison of the proposed method's performance against existing methods based on various metrics such as training and testing accuracy, number of epochs, accuracy plot, loss plot, confusion matrix, and classification matrices. Further, the comparison between the audio, visual, and audiovisual models is discussed.

4.1 Acoustic speech recognition evaluation

The English dataset was trained for 60 epochs and the training accuracy reached 93.67%, as shown in Fig. 5. The test accuracy achieved was 91.53%.

We compared the training and validation accuracies of the proposed model based on the number of epochs used. The results are presented in Fig. 6. The figure indicates the variation in the training and validation accuracy as the number of epochs increases. Additionally, Fig. 7 illustrates the changes in the training and validation losses as the number of epochs increases.

```
val_accuracy: 0.9153
Epoch 195/200
18/18 [==============================] - 0s 18ms/step - loss: 0.2683 - accuracy: 0.9140 - val_loss: 0.2569 -
val_accuracy: 0.9101
Epoch 196/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2551 - accuracy: 0.9052 - val_loss: 0.2800 -
val_accuracy: 0.8942
Epoch 197/200
18/18 [==============================] - 0s 16ms/step - loss: 0.2698 - accuracy: 0.9117 - val_loss: 0.2455 -
val_accuracy: 0.8995
Epoch 198/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2771 - accuracy: 0.8985 - val_loss: 0.2019 -
val_accuracy: 0.9206
Epoch 199/200
18/18 [==============================] - 0s 17ms/step - loss: 0.2879 - accuracy: 0.9038 - val_loss: 0.2432 -
val_accuracy: 0.9206
Epoch 200/200
18/18 [==============================] - 0s 17ms/step - loss: 0.1807 - accuracy: 0.9367 - val_loss: 0.2052 -
val_accuracy: 0.9153
```

Fig. 5. Epochs are trained for an audio model.

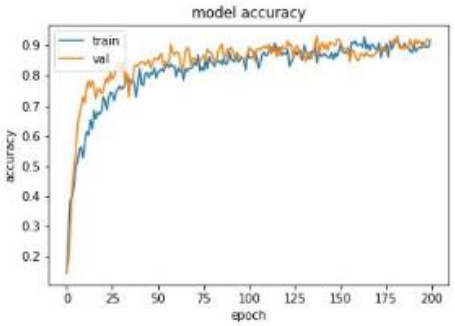


Fig. 6. Audio model accuracy curve: epoch vs. accuracy

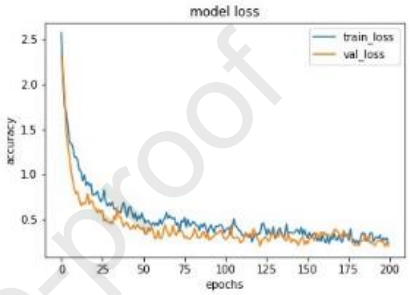


Fig. 7. Epoch vs. precision audio model loss curve

Table 1. Matrix of classification for the acoustic model

Words Name	Precision	Recall	f1-score	support
About	0.85	0.81	0.83	21
Bad	0.88	1.00	0..93	21
Bottle	1.00	0.86	0.92	21
Come	0.90	0.86	0.92	21
Cow	0.78	0.86	0.88	21
Good	0.95	0.86	0.82	21
Pencil	1.00	1.00	1.00	21
Read	1.00	0.90	0.95	21
Where	0.91	1.00	0.95	21
Accuracy	-	-	0.92	189
macro avg.	0.92	0.92	0.92	189
Weighted avg.	0.92	0.92	0.92	189

The number of words categorized into various types is shown in the auditory model confusion matrix, which is trained on English datasets. The precision and misclassification of individual words are shown in Fig. 8. The system correctly identified the first word, “About,” 81% of the time and misclassified it only 19% of the time. “Bad,” the second word correctly predicted that it would be recognized with 100% accuracy without any misclassification. According to the graph, the third word, “Bottle,” was correctly predicted 86% of the time, whereas 14% of the time it was incorrectly classified as “About,” “Bad,” or “Cow.” The fourth word, “Come,” was recognized with an accuracy of 86% and misclassified 14% of the time as “About.” The fifth word, “Cow,” was correctly predicted 86% of the time and misclassified 14% of the time as “About” and “Coming.” The sixth word, “Good,” was correctly classified 95% of the time and incorrectly classified 5% of the time. The seventh word, “Pencil,” was correctly predicted 100% of the time and was recognized with 100% accuracy. The eighth word, “Read,” was predicted with 90% accuracy and incorrectly predicted 10% of the time as “Where.” The ninth word, “Where,” was correctly predicted 100% of the time with 100% accuracy and no misclassification. Table 1

presents a classification report for the audio database. The perceived precision, recall, accuracy, and F1-score of the proposed system were 92%, 92%, 92%, and 92%, respectively.

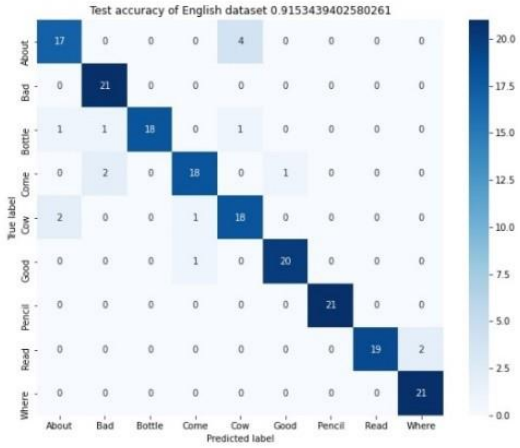


Fig. 8. Matrix of confusion for the acoustic model

4.2 Visual speech recognition evaluation

This section discusses the visual model epoch information, model precision, loss accuracy, confusion matrix, and classification details.

```
Epoch 55/60
18/18 [==============================] - 1s 42ms/step - loss: 0.7268 - accuracy: 0.7414 - val_loss: 1.1511 - val_
accuracy: 0.6667
Epoch 56/60
18/18 [==============================] - 1s 41ms/step - loss: 0.7306 - accuracy: 0.7416 - val_loss: 1.0958 - val_
accuracy: 0.6984
Epoch 57/60
18/18 [==============================] - 1s 41ms/step - loss: 0.6274 - accuracy: 0.7793 - val_loss: 1.2342 - val_
accuracy: 0.7143
Epoch 58/60
18/18 [==============================] - 1s 39ms/step - loss: 0.6977 - accuracy: 0.7549 - val_loss: 1.3743 - val_
accuracy: 0.6720

Epoch 59/60
18/18 [==============================] - 1s 38ms/step - loss: 0.7350 - accuracy: 0.7513 - val_loss: 1.1761 - val_
accuracy: 0.6720
Epoch 60/60
18/18 [==============================] - 1s 38ms/step - loss: 0.6593 - accuracy: 0.7748 - val_loss: 1.0564 - val_
accuracy: 0.7619
```

Fig. 9. Epochs for visual model training

The visual model was trained on the English visual dataset using 60 epochs, achieving a training accuracy of 77.48% and test accuracy of 76.19%, as shown in Fig. 9. Fig. 10 illustrates the variation in training and validation accuracy over multiple epochs for our dataset, whereas Fig. 11 displays the fluctuation in training and validation loss with respect to different epochs.

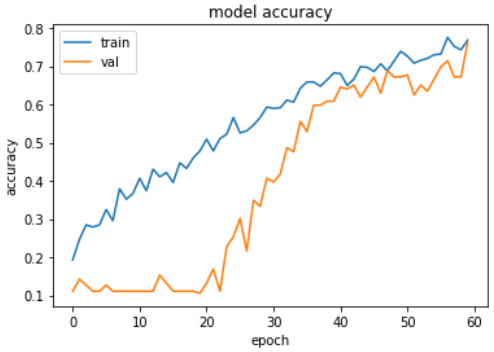


Fig. 10. Epoch vs. accuracy model accuracy curve for visual speech.

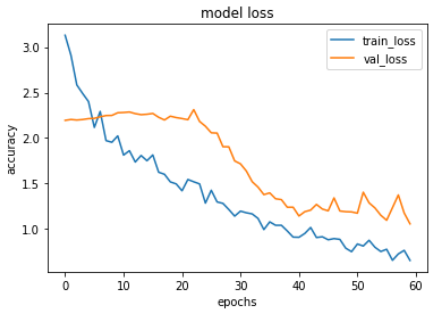


Fig. 11. Epoch vs. accuracy model loss curve for visual speech

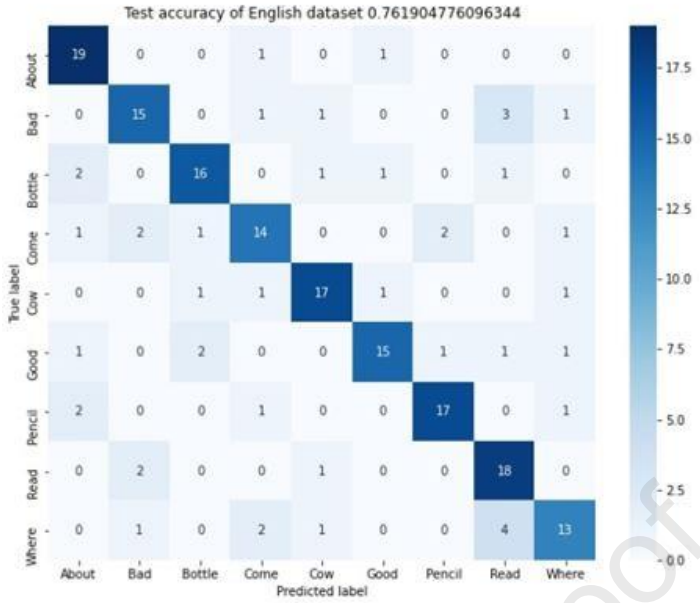


Fig. 12. Matrix of confusion for visual speech

The number of words categorized into various types is shown in the confusion matrix. The English datasets were trained using the confusion matrix of the visual model. The accuracy and misclassification of each word are displayed in Fig. 12. The system correctly identified "About" with 90% accuracy and misclassified "Come" and "Excellent" only 10% of the time. "Bad" was recognized correctly 71% of the time and misclassified 29% of the time as "Come," "Cow," "Read," and "Where." "Bottle" was correctly predicted 76% of the time and incorrectly identified 24% of the time as "About," "Cow," "Excellent," and "Read." "Come" was correctly predicted 66% of the time and incorrectly identified 34% of the time as "About," "Bad," "Bottle," "Pencil," and "Where." "Cow" was correctly predicted 81% of the time and incorrectly predicted 19% of the time as "Bottle," "Coming," and "Good." "Excellent" was correctly identified 71% of the time and misclassified 29% of the time as "About," "Bottle," "Pencil," "Read," or "Where." "Pencil" was correctly identified 81% of the time and misclassified 29% of the time as "About," "Come," and "Where." "Read" was correctly predicted in the graph 86% of the time and incorrectly classified 14% of the time. "Where" was correctly predicted 62% of the time and misclassified 38% of the time as "Bad," "Come," "Cow," and "Read." The classification information for the visual speech dataset is shown in Table 2.

Table 2. Matrix of classification for visual speech

Words Name	Precision	Recall	f1-score	support
About	0.76	0.90	0.83	21
Bad	0.75	0.71	0.73	21
Bottle	0.80	0.86	0.78	21
Come	0.70	0.67	0.68	21
Cow	0.81	0.81	0.81	21
Good	0.83	0.71	0.77	21
Pencil	0.85	0.81	0.83	21
Read	0.67	0.86	0.75	21
Where	0.72	0.62	0.67	21
Accuracy	-	-	0.76	189
macro average	0.77	0.76	0.76	189
Weighted average	0.77	0.76	0.76	189

4.3 Integration of acoustic and visual speech evaluation

In this section, the results of the fusion model are discussed in terms of the number of epochs, model accuracy, accuracy loss, confusion matrix, and classification matrix. English datasets were used to train the audiovisual

speech recognition model for 100 epochs. The fusion model achieved a training accuracy of 94.67% and test accuracy of 91.75% for the English dataset, as shown in Fig. 13. For our dataset, Fig. 14 shows the variation in training and validation accuracies over different epochs. The loss accuracy of the fusion model, shown through the training and validation losses, is shown in Fig. 15.

```
accuracy: 0.8981
Epoch 95/100
17/17 [==============================] - 1s 67ms/step - loss: 0.1211 - accuracy: 0.9520 - val_loss: 0.4226 - val_
accuracy: 0.9029
Epoch 96/100
17/17 [==============================] - 1s 88ms/step - loss: 0.1787 - accuracy: 0.9386 - val_loss: 0.3460 - val_
accuracy: 0.9126
Epoch 97/100
17/17 [==============================] - 2s 95ms/step - loss: 0.1553 - accuracy: 0.9438 - val_loss: 0.3529 - val_
accuracy: 0.9029
Epoch 98/100
17/17 [==============================] - 2s 92ms/step - loss: 0.1661 - accuracy: 0.9470 - val_loss: 0.3456 - val_
accuracy: 0.9126
Epoch 99/100
17/17 [==============================] - 2s 91ms/step - loss: 0.1960 - accuracy: 0.9422 - val_loss: 0.3751 - val_
accuracy: 0.9126
Epoch 100/100
17/17 [==============================] - 2s 93ms/step - loss: 0.2047 - accuracy: 0.9467 - val_loss: 0.3180 - val_
accuracy: 0.9175
```

Fig. 13. Epochs for audiovisual speech model training

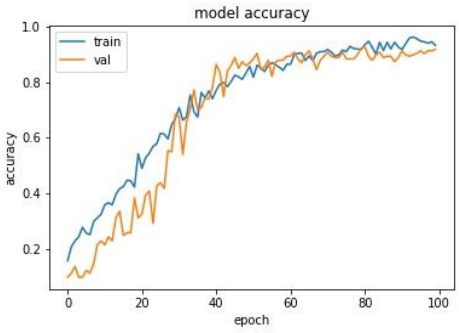


Fig. 14. Epoch vs. accuracy audiovisual speech model accuracy curve

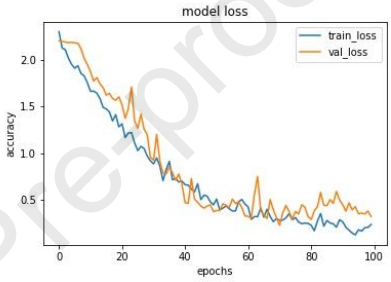


Fig. 15. Epoch vs. accuracy audiovisual speech model loss curve

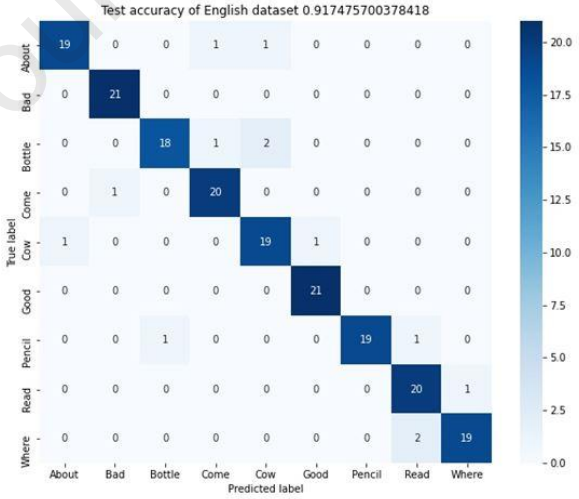


Fig. 16. Confusing speech for audiovisual

The number of words categorized into various types is shown in the confusion matrix. The English datasets were trained using the confusion matrix of the audiovisual model. The accuracy and misclassification of each word are shown in Fig. 16. The system correctly identified “About” with 90% accuracy and misclassified “Come” and “Cow” only 10% of the time. “Bad” was correctly predicted 100% of the time without error. “Bottle” was correctly identified 86% of the time and misclassified 14% of the time. “Come” was recognized correctly 95% of the time and misclassified 5% of the time. Five-word recognition rates for “Cow,” “About,” and “Excellent”

showed a 90% accuracy and 10% misclassification level. “Good” was predicted accurately 100% of the time without error. “Pencil” was correctly classified 90% of the time and incorrectly classified 10% of the time as “Bottle” or “Read.” “Read,” was correctly predicted 95% of the time and incorrectly predicted 5% of the time as “Where.” “Where” was predicted accurately 90% of the time and misclassified 10% of the time as “Read.” The categorization report for the audiovisual speech database is presented in Table 3. The perceived precision, recall, accuracy, and F1-score of the proposed method were 92%, 92%, 92%, and 92%, respectively. Table 4 compares the results of visual-only speech recognition with the suggested outcome, and audiovisual speech recognition with the suggested outcome.

Table 3. The audiovisual speech classification matrix.

Words Name	Precision	Recall	f1-score	support
About	0.85	0.81	0.83	21
Bad	0.88	1.00	0.93	21
Bottle	1.00	0.86	0.92	21
Come	0.90	0.86	0.88	21
Cow	0.78	0.86	0.82	21
Good	0.95	0.95	0.95	21
Pencil	1.00	1.00	1.00	21
Read	1.00	0.90	0.95	21
Where	0.91	1.00	0.95	21
Accuracy	-	-	0.92	189
macro avg.	0.92	0.92	0.92	189
Weighted avg.	0.92	0.92	0.92	189

Table 4. Comparison of the proposed method with current audiovisual and visual methods for speech recognition.

Method	Dataset	Correctness	Word Error Rate
DWT/LDA[18] [v]	Custom AAVC	68.04%	31.96%
VGG16[20] [v]	Customized	75%	25%
PCA [32] [v]	Ryerson Multimeida Lab	75%	25%
LSTM[32] [v]	Customized	75%	25%
LSTM [Proposed]	Customized	76.19%	25%
Decision Fusion [18][AV]	Customized AAVC	76.79%	23.21
PCA LDA [32]	Ryerson Multimeida Lab	82.5%	17.5%
DCNN[32]	Customized	91%	9%
DCNN[Recommended]	Customized	91.74%	8%

4.4 Management implications and discussion

Understanding and sensing the meanings behind what people communicate by simply observing visuals without audio is quite complex. Deep learning models are used to predict what someone is saying using visual information, primarily lip movements. Voice biometric applications, for example, are often used by office and university managers to verify individuals. Physically challenged people, such as those with audio and verbal speech impairments, can use such applications to manage conversations without the involvement of audio language. The audiovisual speech recognition method proposed in this study involves a new dataset to address the computational challenges of CNNs and deep learning.

5. Conclusion and future scope

The audiovisual speech recognition method proposed here involves a new dataset and a feed-forward neural network that addresses the computational challenges of CNNs and deep learning. The architecture includes a 1D CNN model for audio, an LSTM model for visual, and a DCNN for integration. The authors achieved a training accuracy of 94.67% and testing accuracy of 91.75% using a custom dataset. The results demonstrate that combining audio and visual inputs yields the best performance, as evidenced by the superior performance of the proposed method compared with existing methods.

The performance of audiovisual voice recognition for English-language words was examined and the results were compared with those of earlier approaches and implementations. For subsequent research, the authors suggest that employing additional datasets to train various deep learning algorithms and building a dataset from perspectives other than looking directly at the speaker, as well as working with various machine learning and deep learning algorithms for better results. In the future, researchers will create datasets in noisier environments or add artificial noise and train the system using different machine learning and deep learning algorithms. Researchers can implement this in the real world in an unsupervised manner and develop a portable device using a microcontroller and microprocessors.

Conflicts of interest:

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

Availability of data and material:

Data supporting the findings of this study are available from the corresponding author upon request.

Code availability:

The code is available on reasonable request to the corresponding author.

References

[1]. Abderrahim, Mesbah, Aissam, Berrahou, Hicham, Hammouchi et al., 2019. Lip Reading with Hahn Convolutional Neural Networks moments. Image Vis. Comput., 88(Aug), 76–83.

[2]. Abdelwahab, M., and Busso, C., 2019. Active Learning for Speech Emotion Recognition Using Deep Neural Network. 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 1-7.

[3]. Afouras, T., Chung, J., S., Senior A.,et al. 2018 Deep Audio-visual Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), pp. 8717-8727.

[4]. Afouras, T., Chung, J., S., and Zisserman A., 2018. LRS3-TED: a largescale dataset for visual speech recognition. Available at http://arxiv.org/abs/1809.00496.

[5]. Agarwal, S., Das, D., and Bhowmick, B., 2021. Realistic lip animation from speech for unseen subjects using few-shot cross-modal learning. 2020 28th European Signal Processing Conference (EUSIPCO), IEEE, pp. 690-694.

[6]. Befkadu, Belete, Frew, 2019. Audio-Visual Speech Recognition using Lip Movement for Amharic Language. Int. J. Eng. Res., 8(8), pp. 594–604.

[7]. Cornejo, J., Y., R., and Pedrini, H., 2019. Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC),IEEE, pp. 3396–3402.

[8]. Debnath, S., Roy, P., 2021. Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition. SIViP 15, 25–32.

[9]. Dupont, S. and Luettin, J.,2000. Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia, 2(3), pp. 141-151.

[10]. Feng, W., Guan, N., Li Y.et al. 201. Audiovisual speech recognition with multimodal recurrent neural networks. 2017 International Joint Conference on Neural Networks (IJCNN), pp. 681–688.

[11]. George, Sterpu, Naomi, Harte, 2022. Taris: An online speech recognition framework with sequence to sequence neural networks for both audio-only and audio-visual speech. Computer Speech & Language, 74(July), 101349.

[12]. Goh, Y., H., Lau, K., X., and Lee, Y., K., 2019. Audio-Visual Speech Recognition System Using Recurrent Neural Network. 2019 4th International Conference on Information Technology (InCIT),IEEE, pp. 38–43.

[13]. Heckmann, M., Berthommier, F., Kroschel, K., 2002. Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition. EURASIP J. Adv. Signal Process,11,1260-1273.

[14]. Jian, Wu, Yong, Xu, and Shi-Xiong, Zhang, et al., Time Domain Audio-Visual Speech Separation. 2019 IEEE Autom. Speech Recognit. Underst. Work. ASRU. IEEE, pp. 667–673.

[15]. Jang, D., W., Kim, H., I., Je, C., et al., 2019. Lip reading using committee networks with two different types of concatenated frame images. IEEE Access, 7, pp. 90125–9013.

[16]. Jadczyk, T., 2018. Audio-visual speech-processing system for Polish applicable to human-computer interaction. Comput. Sci.,19(1), pp. 41–64.

[17]. Ken, W., Grant, Brian, E., Walden, Philip, F., Seitz., 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am, 103 (5): 2677–2690.

[18]. Li, X., Neil, D., Delbruck, T., et al., 2019. Lip reading deep network exploiting multi-modal spiking visual and auditory sensors. 2019 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, pp. 1-5.

[19]. Makino, T., Liao, H., Assael, Y., et al., 2019. Recurrent Neural Network Transducer For Audio-Visual Speech Recognition. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, pp. 905-912.

[20]. Martinez, B., Ma, P., Petridis, S., et al., 2020. Lipreading Using Temporal Convolutional Networks. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., IEEE, pp. 6319–6323.

[21]. Meutzner, H., Ma, N., Nickel, R.,et al., 2017. Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.,IEEE, pp. 5320–5324.

[22]. Nadeemhashmi, S., Gupta, H., Mittal, D.et al., 2018. A Lip Reading Model Using CNN with Batch Normalization. 2018 Eleventh International Conference on Contemporary Computing (IC3), IEEE, pp. 2–4.

[23]. Nefian, A.V., Liang, L., Pi, X. et al., 2002. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. EURASIP J. Adv. Signal Process. 2002, 783042.

[24]. Noda, K., Yamaguchi, Y., Nakadai, K., et al., 2015.Audio-visual speech recognition using deep learning. Appl Intell 42, 722–737.

[25]. Petridis S, Stafylakis T, Ma Pet al., 2018. End-to-End Audiovisual Speech Recognition. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp. 6548–6552.

[26]. Petridis, S., Stafylakis, T., Ma, P., et al., 2019.Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. 2018 IEEE Spoken Language Technology Workshop (SLT),IEEE, pp. 513–520.

[27]. Santos, T., I., and Abel, A., 2019. Using Feature Visualization for Explaining Deep Learning Models in Visual Speech. 2019 4th IEEE International Conference on Big Data Analytics, (ICBDA),IEEE, pp. 231–235.

[28]. Shashidhar, R., Patilkulkarni, S., 2021.Visual speech recognition for small scale dataset using VGG16 convolution neural network. Multimed Tools Appl 80, 28941–28952.

[29]. Shashidhar, R., Patilkulkarni, S., & Puneeth, S., B., 2022. Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int. j. inf. tecnol. 14, 3425–3436.

[30]. Shashidhar, R., Patilkulkarni, S., 2022. Audiovisual speech recognition for Kannada language using feed forward neural network. Neural Comput & Applic 34, 15603–15615.

[31]. Song, Q., Sun, B., and Li, S., Multimodal Sparse Transformer Network for Audio-Visual Speech Recognition. IEEE Transactions on Neural Networks and Learning Systems.(April), 1-11.

[32]. Sooraj, V., Hardhik, M., Murthy, N., S., Sandesh, C., & Shashidhar, R., 2020. Lip-Reading Techniques: A Review. International Journal of Scientific Technology and Research, 9(2), pp. 4378–4383.

[33]. Stafylakis, T., and Tzimiropoulos, G., 2017. Combining residual networks with LSTMs for lipreading. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2017(Augu), pp. 3652–3656.

[34]. Tao, F., and Busso, C., 2018. Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Process., 26(7), pp. 1286–1298.

[35]. Tao, F., and Busso, C., 2021. End-to-End Audiovisual Speech Recognition System with Multitask Learning. IEEE Trans. Multimed., 23, pp. 1–11.

[36]. Tan, K., Xu, Y., S., Zhang, X., et al. 2020. Audio-Visual Speech Separation and Dereverberation with a Two-Stage Multimodal Network. IEEE J. Sel. Top. Signal Process., 14(3), pp. 542–553.

[37]. Tao, F., and Busso, C., 2018. Audiovisual speech activity detection with advanced long short-term memory. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2018(Sept), pp. 1244–1248.

[38]. Wand, M., and Vu, N., T., 2018. Investigations On End-To-End Audiovisual Fusion. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),IEEE, pp. 3041-3045.

[39]. Wei, J., Yang, F., Zhang, J., et al., 2018.Three-dimensional joint geometric-physiologic feature for lip-reading. 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE(Nov), pp. 1007–1012.

[40]. Yu W., Zeiler S., and Kolossa D., 2021. Multimodal integration for large-vocabulary audio-visual speech recognition. Eur. Signal Process. Conf., 2021(1), pp. 341–345.

[41]. Yuan, Y., Tian, C., and Lu, X., 2018. Auxiliary Loss Multimodal GRU Model in Audio-Visual Speech Recognition. IEEE Access, 6, pp. 5573–5583.