Sentences Prediction Based on Automatic Lip-Reading Detection with Deep Learning Convolutional Neural Networks Using Video-Based Features

Khalid Mahboob(✉), Hafsa Nizami, Fayyaz Ali, and Farrukh Alvi

Department of Software Engineering, Sir Syed University of Engineering and Technology,
Karachi, Pakistan
kmahboob@ssuet.edu.pk

Abstract. Lip-reading is the process of deciphering text from a speaker's visual interpretation of facial, lip, and mouth movements without using audio. The challenge is traditionally divided into two stages: creating or learning visual characteristics and prediction. End-to-end techniques for deep lip-reading have been popular in recent years. Existing work on end-to-end models, on the other hand, only does word classification rather than sentence-level sequence prediction. Longer words improve human lip-reading ability, suggesting the relevance of characteristics that capture the temporal context in an inconsistent communication channel. In this study, an end-to-end model based on deep learning convolutional neural network shave been employed to develop an automated lip-reading system that uses a re-current network spatiotemporal convolutions, and the connectionist temporal classification loss to translate a variable-length series of video frames to text. The accuracy of the trained lip-reading process in predicting sentences was evaluated using video-based features.

Keywords: Lip-reading · Convolutional neural networks · Model

1 Introduction

With their remarkable potential to extract patterns and detect trends from difficult or imprecise data, convolutional neural networks may be used to extract patterns and detect too complex trends for humans or other computer approaches to discover. A trained convolutional neural network can be considered an expert in the category of data it is assigned to examine [1]. LIP-READING, one of the simplest techniques to identify speech, uses a convolutional neural network. It is a relatively new approach for voice recognition that is frequently used. A lip-reading method can be described that takes into account both the shape of the lips and the intensity of the mouth area [2].

A. Mohamed et al. (Eds.): SCDS 2021, CCIS 1489, pp. 42–53, 2021.
https://doi.org/10.1007/978-981-16-7334-4_4

This is also a technique for identifying the contents of speeches using just visual data rather than audible data. It is intended to be applied in situations where speech recognition technology is difficult to use, such as in high-noise conditions to gather audio data in a public area where producing a voice is difficult, and to assist people with hearing and speech impairments in communicating [3]. Despite the fact that lip-reading has been explored for decades, there are several obstacles in lip-reading tasks, including the identification target: singular sound, separated word, constant word, and phrase, accessible modalities: audible or only-visual, face orientation, and language [4].

As discussed earlier, lip-reading is the process of extracting visual speech characteristics from a person's lips. The inner and outer lip contours carry the greatest visual speech information, but it has also been discovered that information regarding the appearance of teeth and tongue gives crucial speech signals. The disparity in pronunciation and relative accent of words and phrases is due to the variety of languages spoken throughout the world. Developing a software program that interprets spoken words automatically and properly based simply on the speaker's visual lip movement becomes extremely difficult [5].

Lip-reading methods have improved significantly as a result of the advent of deep learning. The 3D-CNN (usually a 3D convolutional layer tracked by a deep 2D Convolutional Network) has recently gained popularity as a preferred front-end option. CNN, which comprises alternating convolutional and pooling layers, has been an effective model for extracting visual information for image identification and classification applications. The inner product of the linear filter and the receptive field is computed by the convolutional layers, which are then followed by a non-linear activation function (e.g. ReLu, sigmoid, tanh, etc.) [6].

The CNN's output features are input into two Bidirectional Gated Recurrent Network (GRU) layers, which are then followed by a linear transformation and a softmax over the corpus through each time step (which is a character-based interpretation in this situation). A Connectionist Temporal Classification (CTC) network with a softmax output layer with many labels in the vocabulary with one blank character is used to train this end-to-end model. The CTC determines the probability of all possible string combinations [6, 7].

The purpose of this study is to propose a system that will help disabled people that cannot hear properly so that they can get the lip-synch through CNN that will generate text for them so that they can have their conversation properly. The system features a deep learning model for lip-reading detection that will consider the video as an input and generate text as the description. The main objective of this study is to predict sentence-level sequences based on automatic lip-reading detection using a convolutional neural network with video-based features [14]. Several model hyperparameters, such as picture size, filter size, number of filters, activation function, types, and layer layout, number of layers, were optimized to find the best architecture for CNN [15]. For imaging data classification, CNN is recognized as more effective and efficient [16].

The paper is organized as follows: Sect. 2 offered a comprehensive literature review. The pre-processing and methodology are stated in Sect. 3. Section 4 offered brief results and discussion on an exploration followed by the last section i.e., Sect. 5 with a conclusion.

2 Literature Review

The visual recognition system proposed in this study reads the gestures made by the lips of the user and tries to guess the word that is being said by the user. There is a scarcity of such tools which could be used in the generation of the content using visual lip gestures which could be helpful for people with speech and auditory disabilities. It is a difficult task to develop a system that could read the lip expression of the user and translate it in a language that the disabled person best understands.

Vaishali A. Kherdekar et al. [1] have presented the experiment for speech recognition of mathematical phrases which is useful to people with disabilities. The CNN model is used to increase the recognition accuracy. The researchers have chosen 17 mathematical terms that are often used in mathematical expressions. Because of its speed, the Rectified Linear Unit Activation Function is utilized to train CNN. For Adam and Adagrad optimizers, this study examines the model for MFCC and Delta MFCC features. The results demonstrate that for both Adam and Adagrad optimizers, Delta MFCC provides an accuracy of 83.33 percent. It shows that Delta MFCC is superior to MFCC in terms of outcomes. Adagrad with Delta MFCC trains the model earlier than Adam, according to the results.

Densely Connected Temporal Convolutional Network (DC-TCN) was introduced by Pingchuan Mal et al. [2] to recognize people's isolated words. To capture more strong temporal characteristics, they added dense connections to the network. To improve the model's classification power, the technique employs the Squeeze and-Excitation block, a lightweight attention mechanism.

Tasuya Shirakata et al. [3] used a technique for identifying what individuals are saying using just visual data rather than auditory data. It is useful for handicapped persons and patients. It refers to the location where the noise problem arises. In the lip reading approach, it incorporates facial expression elements such as expression-based feature and action-based unit feature. Experiments were carried out in the OuluVS, CUAVE, and CENSREC-1-AV public databases.

Karan Shrestha [5] used lip-reading techniques to detect the visual interpretation based on the movement of the face, mouth, and lips without the need for audible data. To predict real words, CNN architectures are implemented. To train the robust lip-reading system, the entire dataset is utilized in the LRW dataset. In addition, the lightweight architecture may be advanced.

Denis Ivanko et al. [7] analyzed the impact of several audiovisual fusion methods on word recognition accuracy rates in English and Russian (on GRID and HAVRUS corpora, respectively). GMMCHMM, DNN-HMM, and end-to-end techniques were examined as tree audiovisual modalities integration strategies. The conventional GMM-CHMM method produced the best recognition results. The GRID dataset has demonstrated that NN-based approaches are superior to the old GMM-CHMM approach to audiovisual speech recognition in all experiments.

M.Rahman et al. [8] suggest an automated speech recognition system built on Support Vector Machine (SVM) along with the help of dynamic time warping (DTW) for Bengali speaking individuals. The data was collected from 40 Bangla-speaking individuals for five different Bengali words with minimal noise environment and acoustics with high amplitudes. DTW was used after determining the feature vectors. A reliable model was proposed that was tested with 12 speakers with the recognition of 86.08%. The only apparent setback for their proposed system was that this system was specifically developed for Bengali speech recognition and cannot be generalized.

Garg et al. [9] proposed a combination of the CNN and LSTM deep learning methods and applied it to the lip gestures data collection problem. CNN in this scenario was used for feature extraction (reading lip gestures) and LSTM for classification.

Because CNN contains characteristics such as dimensionality reduction and parameter sharing, it outperforms other machine learning algorithms. The number of parameters in CNN is decreased as a result of parameter sharing, and therefore the computations are also reduced. Further, when CNN's feature extraction grows deeper, it delivers improved image identification (encompasses additional layers). That is why CNN is selected for automatic lip-reading in this study.

3 Pre-processing and Methodology

The system we developed based on CNN, takes a video as an input and predicts the speaker's lip-synch which is different from the study in [10] as they proposed the lip images feature. Because the model is trained on the grid dataset, it will predict data from the grid dataset. The model has three spatiotemporal convolutions, channel-wise dropout, and spatial max-pooling layers, as well as two Bi-GRUs layers, all of which include rectified linear unit (ReLU) activation functions and a softmax activation function for sequence classification. The overall flow of the system processing is shown in Fig. 1.

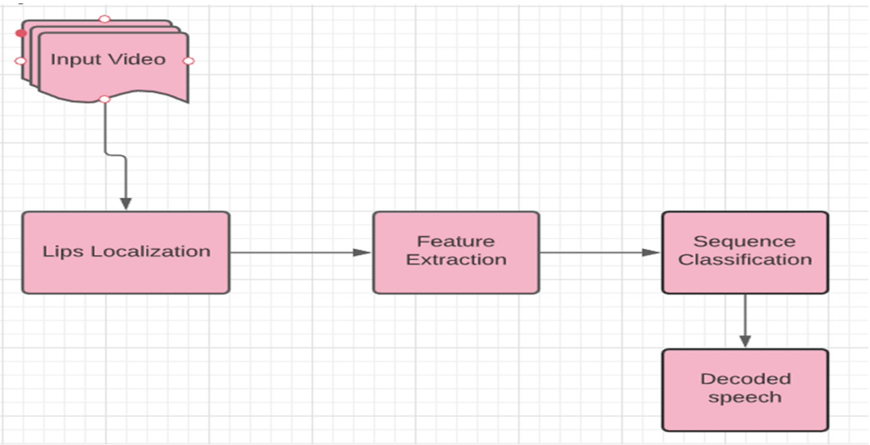

Fig. 1. Overall flow of the process

The desktop PC on which the model runs is the hardware component; the operating system is the software component; the operating system we chose is Linux (Ubuntu). And the video should be in mpg format, according to the model. The user may produce a video on any device, convert it to mpg format, and then execute a program on Linux. The system is not user-friendly since the user must run a command to make a prediction, and the model does not operate in real-time.

The data is in the form of a video file with the extension.mpg, which will be utilized as an input. The video will be processed by the model's learned weights, and the model will predict the speaker's phrase based on the grid data set corpus. Pre-processing is a crucial stage in any deep learning or machine learning system. The mouth has been cropped to a size of 100 × 50 pixels every frame as part of the preprocessing. To get a zero mean and unit variance, the RGB channels normalized throughout the whole training set. To minimize overfitting, we have added basic transformations to the dataset.

Keras comes pre-loaded with many algorithms designed particularly for deep learning. Different layers exist in models. We have used an input layer that accepts input in the form of img _frames, img_width, img_height, and img_channel if the image data format is Theano will take precedence. For padding zeros in three dimensions, we have used the ZeroPadding3D layer. Because the data is in video format, we have used Conv3D layers and the rectified linear unit as the activation function, which makes a positive input equal and a negative input zero, and is the usual activation function for deep learning applications. The Batch Normalization layer has been used to normalize the input [11].

We have utilized Spatial Dropout3D layers to keep the model from being too tight. We have also developed MaxPooling3D layers, which will employ three-dimensional max pooling to discover the most features. As an input to the wrapper layer, we applied Time Distributed wrapper layers with flattening layers. We have also used BiDirectional wrapper layers with GRU layers as wrapper inputs. BiDirectional wrapper layers help GRU layers train quicker and more efficiently. For sequence-wise classification, we employed a dense layer with a softmax activation function.

The output data is a video with sentences predicted in it, with the predicted sentence written on the console and as subtitles. The grid corpus data set is implemented for training the model, which consists of 34 speakers telling 1000 words each. NumPy, Keras, and Tensor Flow are some of the libraries, used for face detection. Python is the programming language that is used here, and all of the libraries are written in Python. The entire system flow architecture is illustrated in Fig. 2.

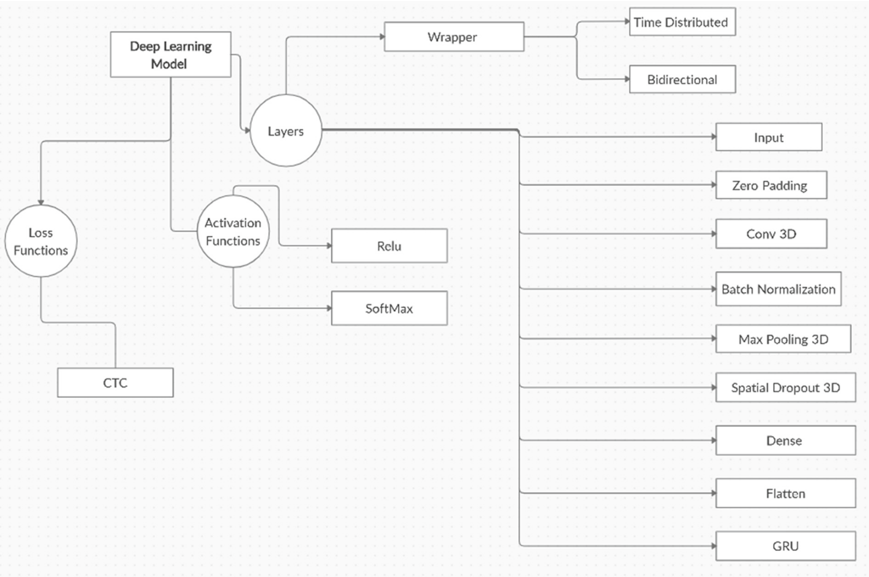

Fig. 2. System flow architecture

4 Results and Discussion

As a dataset, the GRID corpus was employed. It includes videos of each of the 34 speakers speaking 1000 sentences (17 male, 16 female). The sentences are structured in the following way: command + color + preposition + letter + digit + adverb. Bin, lay, place, and the set is among the commands, and the colors are blue, green, red, and white. At, by, in, and with are the five prepositions used, and all Latin letters except W are used. The digits range from 0 to 9, and the adjectives are: again, now, please, soon. 'bin white at f zero again' is an example statement from the dataset. Transcription of the words said, as well as information on when each word is pronounced, is included with each video [12]. The architecture of the proposed convolutional neural network (CNN) is shown in Fig. 3.

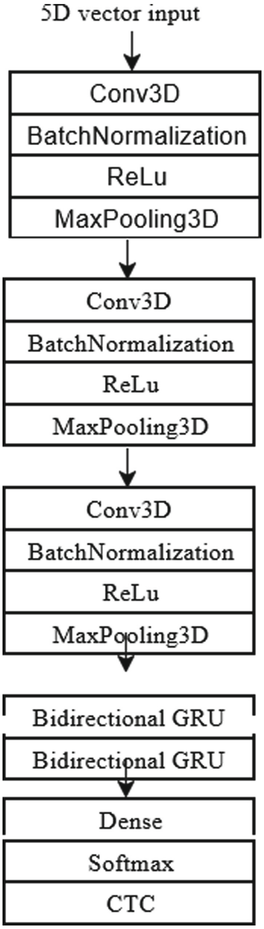

Fig. 3. The architecture of the proposed convolutional neural network (CNN)

Each video in the dataset is divided into 75 frames to prepare it for training. Each frame is then evaluated with the Face Recognition API for Python, which is based on dlib's deep learning-based face recognition. It calculates the (x, y) coordinates of 67 facial landmarks, such as the eyes, nose, mouth, and chin, as illustrated in Fig. 4. The speaker's mouth coordinates are recorded, while the rest are deleted. The mouth is then normalized to the smallest (x, y) coordinates feasible [13].

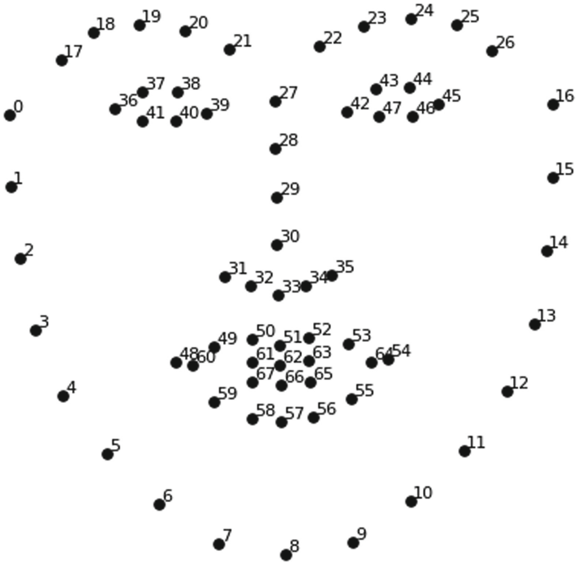

Fig. 4. The 67 Facial landmarks identified with dlib's facial recognition

Because letters may be missing or due to other problems, the CTC layer's output frequently requires processing. The output was run through a spelling algorithm that searched up words in the dictionary and picked the one with the shortest Levenshtein distance, or the term with the most occurrences in the training labels if there were many possibilities. With kenLM and ARPA-format, a 3-g model was created and implemented in the model assessment procedure. Following the spelling process, the CTC output was routed to the LM. Speech recognition accuracy can be improved by using a mix of CTC and an N-gram LM. Different speakers lip-synching a sentence without audio is depicted in Fig. 5.

The ARPA model was created using a text file containing all GRID corpus sentences. The probability of a word was provided in log10, and a term that was not in the lexicon was assigned an absolute penalty of -100 probability because the model did not know the correct probability of that word. The sequence of sentences with the accuracy achieved is shown in Table 1 and Fig. 6. It is important to note that from the experiment results, Letter $\{A - Z/a - z\}$ has achieved the highest accuracy i.e. 79% whereas Digit {0–9} has achieved the least accuracy i.e. 44.5%.

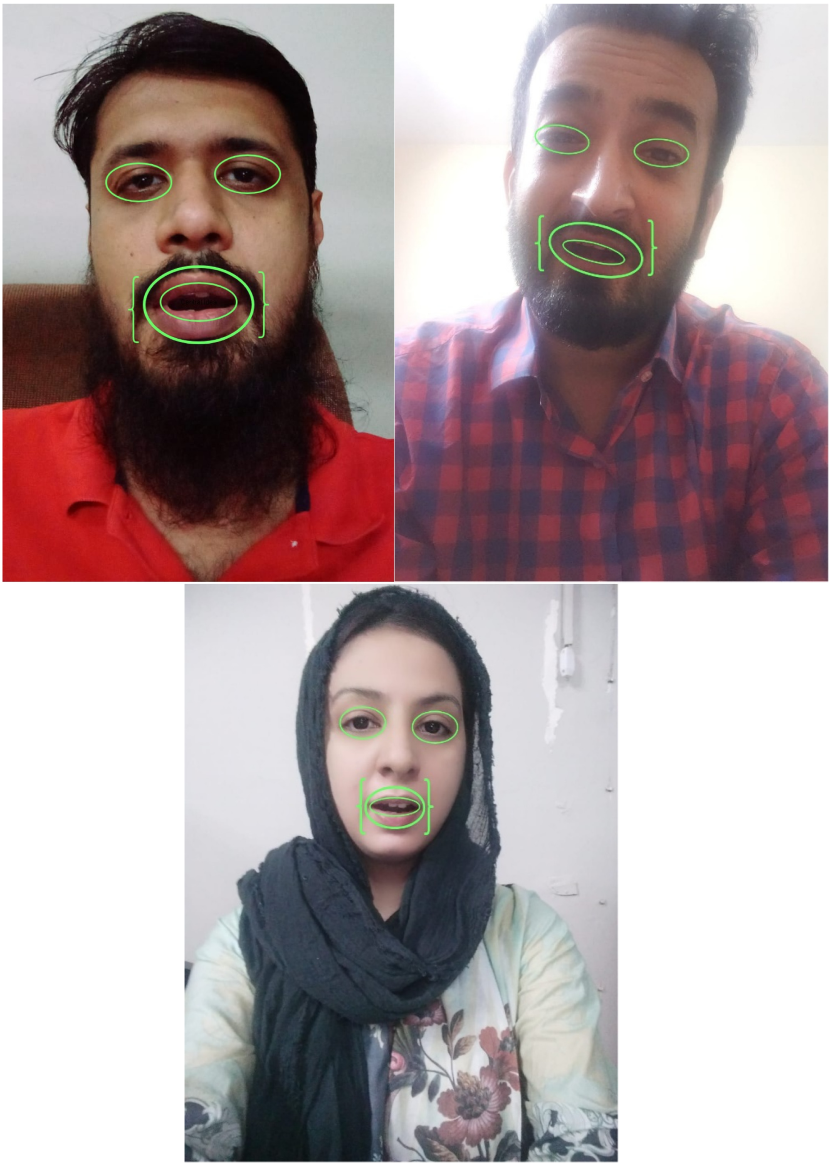

Fig. 5. Three speakers lip-synching a sentence

Table 1. Structure of the sentences with their Accuracies

Sentences	Accuracy
Command	74.00%
Color	64.00%
Preposition	66.15%
Letter	79.00%
Digit	44.50%
Adverb	76.00%

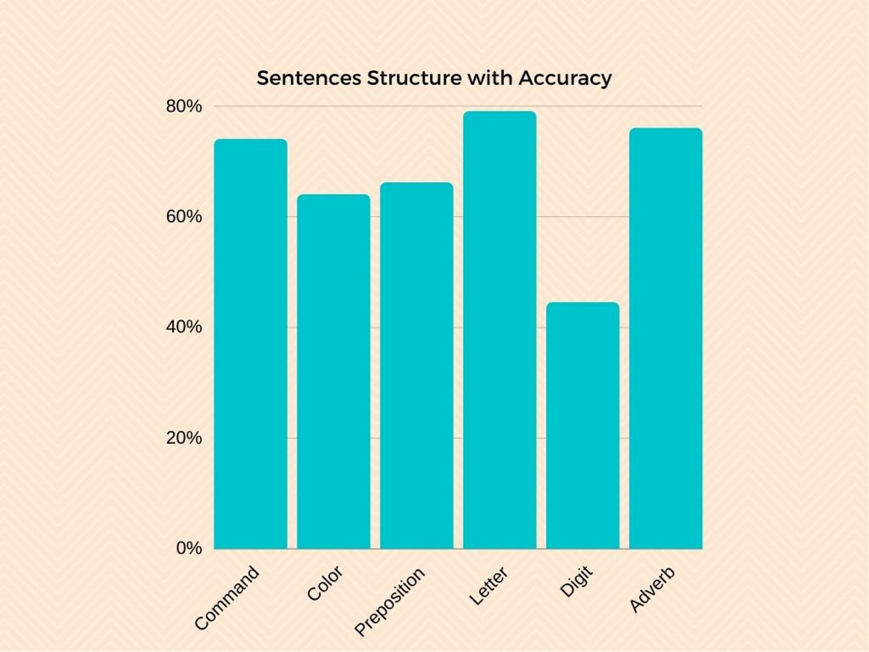

Fig. 6. Comparison of the sequence of the sentences prediction with Accuracy

5 Conclusion

The major goal of this study was to look into the possibility of employing facial landmarks based on deep learning convolutional neural networks to train a model for lip-reading with video-based features. The facial landmark representation proposed in a study captured sufficient information about the speaker. However, it appears that this depiction does not include all of the essential characteristics for lip-reading that may be identified in videos. Except for the lips, the present technique for extracting the speaker's facial features is unable to recognize any aspects of the mouth. The capacity to track teeth and tongue, based on the results and discussion, would most certainly improve the model's ability to discriminate between visages and therefore improve accuracy.

Because various resolutions result in varying distances between the coordinates, normalizing the distance between the coordinates would reduce the disparities between videos and therefore improve the model's generalization ability. Additional specialized landmark extractions perhaps capture more critical characteristics, such as teeth and tongue, to increase the accuracy of the suggested model.

References

1. Kherdekar, V.A., Id, E., Naik, S.A.: Convolution neural network model for recognition of speech for words used in mathematical expression. Turk. J. Comput. Math. Educ. **12**(6), 4034–4042 (2021)
2. Ma, P., Wang, Y., Shen, J., Petridis, S., Pantic, M.: Lip-reading with densely connected temporal convolutional networks, pp. 2857–2866 (2021)
3. Shirakata, T., Saitoh, T.: Lip reading using facial expression features. Int. J. Comput. Vis. Signal Process. **1**(1), 9–15 (2020)
4. Ozcan, T., Basturk, A.: Lip reading using convolutional neural networks with and without pre-trained models. Balk J. Electr. Comput. Eng. (July), 195–201 (2019)
5. Shrestha, K.: Lip reading using neural network and deep learning
6. Fernandez-Lopez, A., Sukno, F.M.: Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput. [Internet] **78**, 53–72 (2018). https://doi.org/10.1016/j.imavis.2018.07.002
7. Ivanko, D., Ryumin, D., Karpov, A.: An experimental analysis of different approaches to audio–visual speech recognition and lip-reading. Smart Innov. Syst. Technol. **187**, 197–209 (2021)
8. Rahman, M.M., Roy Dipta, D., Hasan, M.M.: Dynamic time warping assisted SVM classifier for bangla speech recognition. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (2018). https://doi.org/10.1109/ic4me2.2018.8465640
9. Garg, A., Noyola, J.: Lip reading using CNN and LSTM. In: Proceedings of the 30th IEEE Conference on Computer Vision Pattern Recognition, CVPR 2017, January 2017, p. 3450 (2017)
10. Li, Y., Takashima, Y., Takiguchi, T., Ariki, Y.: Lip reading using a dynamic feature of lip images and convolutional neural networks. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) (2016). https://doi.org/10.1109/icis.2016.7550888
11. Rahmani, M.H., Almasganj, F.: Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. In: 3rd International Conference on Pattern Recognition and Image Analysis, IPRIA 2017, April 2017, pp. 195–199 (2017)
12. Adeel, A., Gogate, M., Hussain, A., Whitmer, W.M.: Lip-reading driven deep learning approach for speech enhancement. IEEE Trans. Emerg. Top Comput. Intell. **5**(3), 481–490 (2019)
13. Vakhshiteh, F., Almasganj, F.: Lip-reading via deep neural network using appearance-based visual features. In: 2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME) (2017). https://doi.org/10.1109/icbme.2017.8430230
14. Lu, Y., Li, H.: Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Appl. Sci. **9**(8) (2019)

15. Jameel, S.M., Hashmani, M.A., Rehman, M., Budiman, A.: Adaptive CNN ensemble for complex multispectral image analysis. Complexity (2020). https://doi.org/10.1155/2020/8361989
16. Jameel, S.M., Hashmani, M.A., Alhussain, H., Rehman, M., Budiman, A.: An optimized deep convolutional neural network architecture for concept drifted image classification. In: Bi, Y., Bhatia, R., Kapoor, S. (eds.) IntelliSys 2019. AISC, vol. 1037, pp. 932–942. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-29516-5_70