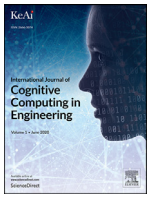
International Journal of Cognitive Computing in Engineering 3 (2022) 24–30

Contents lists available at ScienceDirect

International Journal of Cognitive Computing in Engineering

journal homepage: https://www.keaipublishing.com/en/journals/international-journal-of-cognitive-computing-in-engineering/



Deep learning based assistive technology on audio visual speech recognition for hearing impaired

L Ashok Kumar[a,*], D Karthika Renuka[b], S Lovelyn Rose[c], M C Shunmuga priya[d], I Made Wartana[e]

[a] Department of Electrical and Electronics Engineering, PSG College Of Technology, Coimbatore, India
[b] Department of Information Technology, PSG College Of Technology, Coimbatore, India
[c] Department of Computer Science and Engineering, PSG College Of Technology, Coimbatore, India
[d] Department of Information Technology, PSG College Of Technology, Coimbatore, India
[e] Department of Electrical Engineering, National Institute of Technology (ITN), India

ARTICLE INFO

Keywords:
Speech recognition
Deep learning
Lip reading
Visual speech recognition
RNN-GRU
CNN

ABSTRACT

Assistive technology would be an immense benefit for hearing impaired people by using Audio Visual Speech Recognition (AVSR). Around 466 million people worldwide suffer from hearing loss. Hearing impaired student rely on lip reading for understanding the speech. Lack of trained sign language facilitators and high cost of assistive devices are some of the major challenges faced by hearing impaired students. In this work, we have identified a visual speech recognition technique using cutting edge deep learning models. Moreover, the existing VSR techniques are erroneous. Hence to address the gaps identified, we propose a novel technique by fusion the results from audio and visual speech. This study proposes a new deep learning based audio visual speech recognition model for efficient lip reading. In this paper, an effort has been made to improve the performance of the system significantly by achieving a lowered word error rate of about 6.59% for ASR system and accuracy of about 95% using lip reading model.

1. Introduction

The Speech recognition system is an interdisciplinary field involves natural language processing, signal processing and artificial intelligence. Speech is a continuous sound signal with a sequence of phonemes and the fundamental mode of human interaction whereas hearing impaired people recognize words spoken by reading a lip. The current scenario using assistive technology for hearing impaired students is given in Fig. 1. Some of the widely used practice by hearing impaired students for learning is sign language. Challenges are listed in the perspective of facilitator and student.

Teachers perspective

- lack of trained sign language teachers
- lack of awareness in new technologies.

Students perspective

- Lack of user friendly devices.
- Difficulty in interpretation.
- High Cost Assistive Devices.
- Lack of e-learning contents with subtitles.

To mitigate the difficulties in the current assistive technology scenario in this paper we have proposed audio visual speech recognition for understanding lip reading in a more accurate way.

Audio speech recognition is an automatic process of converting audio features to text. Librispeech (Panayotov, Chen, Povey & Khudanpur, 2015), Timit (Lopes & Perdig˜ao, 2011) are the widely used dataset for automatic speech recognition. Deep Speech (Amodei, Ananthanarayanan, Anubhai, Bai & Battenberg, 2016), LAS (Chan, Jaitly, Le & Vinyals, 2016), Wav2Letter (Collobert, Puhrsch & Synnaeve, 2016) are the ASR architecture using deep learning algorithms with higher recognition performance. In recent year visual Speech recognition system plays a vital role in speech recognition system because it does not require acoustic environment. A Visual Speech Recognition system is an automatic process of detecting spoken words by tracking the speaker's lip movement. This technology provides an alternative way of communication (i.e. Visual communication) for people with hearing impaired problem. Visual Speech Recognition (VSR) plays a vital role in Audio Visual Speech Recognition System (AVSR) because it improves the performance of audio speech recognition system. Nowadays VSR systems are implemented in noisy outdoor environments like driving a car or speaking on the mobile phone. The approaches in VSR system are the

* Corresponding author.
E-mail address: askipsg@gmail.com (L.A. Kumar).

https://doi.org/10.1016/j.ijcce.2022.01.003
Received 4 January 2021; Received in revised form 20 January 2022; Accepted 20 January 2022
Available online 22 January 2022

Fig. 1. Current scenario using assistive technology for hearing impaired students.

traditional statistical and machine learning approaches and the deep learning approach. The word error rates are high in traditional approach. So, researchers have developed a new approach using deep learning to reduce word error rate. Large scale visual speech recognition and lipnet architecture are the popular VSR deep learning architecture. Deep learning system offers higher recognition accuracy and lower word error rate than traditional approaches.

The main contribution of this paper is listed as follows

- We have trained a RNN-GRU based speech to text model and CNN based visual speech to text model.
- We have developed a novel decision level fusion of audio visual speech recognition technique and the results are compared to the baseline system in the perspective of WER.

The content of this paper is organized as follows: a summary of related work in the field of speech recognition is presented in Section 2, followed by the methodology in Section 3 and result analysis in Section 4.

2. Related work

Lip reading is a method of detecting spoken words by observing movement of speaker lips. Visual Speech Recognition is an automatic process of lip reading. Phonemes are the fundamental linguistic unit, while Visemes are the basic visual unit used in lip reading systems. Yet hearing impaired people have trouble identifying the spoken words for a long time (Easton & Basala, 1982) (Fisher, 1968). Petagan et al. (Petajan, Bischoff, Bodoff & Brooke, 1988) introduces the first VSR method, using the height-weight ratio. Researchers are thus focused on the automatic lip reading system known as visual speech recognition (Torfi, Iranmanesh, Nasrabadi & Dawson, 2017) (Vakhshiteh, Almasganj & Nickabadi, 2018). The VSR program was designed previously, with the following features: mutual knowledge, quality features, tongue appearance, teeth appearance (Heckmann, Savariaux, Berthommier & Fr´ed´eric, 2002). The machine learning classifier (Thabet et al., 2018) like Support vector Machine (Gordan, Kotropoulos & Pitas, 2002) (Frolov & Sadykhov, 2009), hidden Markov models (Puviarasan & Palanivel, 2011) is used for lip image classification. Cutting edge deep learning models like convolution neural network (CNN) is been widely used technique in various applications such as fallot recognition, medical applications (Wang et al., 2021) (Zhang, Satapathy, Guttery, Gorriz & Wang, 2021). Recent research works are focused on deep learning algorithms (Alothmany, Boston, Li, Shaiman & Durrant, 2010) (Feng, Guan, Li, Zhang & Luo, 2017) which provide higher accuracy of recognition and a lower rate of word error compared to traditional machine learning algorithms in speech analytic applications. In 2017, Oxford University developed the Deep Learning Architecture lipnet (Assael, Shillingford, Whiteson & Freitas, 2016) and liptype (Pandey & Arif, 2021) an enhanced version of lipnet.

ASR system map variable length audio signal to variable length sequence of words. Although several traditional methods have been identified for speech recognition system using hidden markov model (HMM), Gaussian markov Model (GMM) the word error rate seems to be higher. HMM is a statistical method uses sequence of states suitable for handling temporal nature and variable length sequence of audio signals. Zeghidour et al. (Zeghidour et al., 2018) exploited convolutional language model which is trained over wall street journal dataset, librispeech dataset and inferred that noise condition impacts the recognition model by increasing the word error rate. S. Watanabe et al. (Watanabe, Hori, Kim, Hershey & Hayashi, 2017) introduces a hybrid model combines

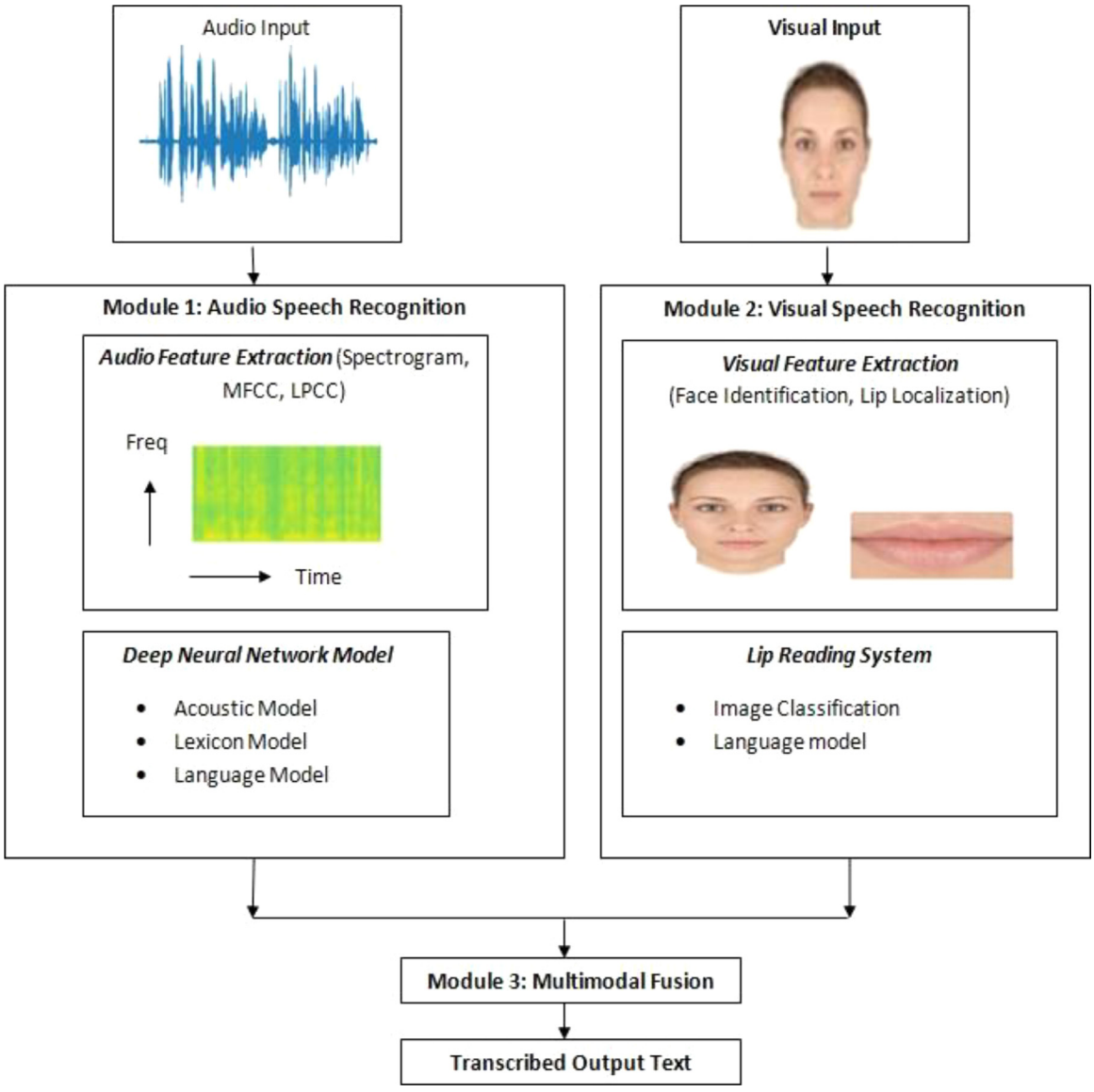


Fig. 2. Architecture of ASR.

CTC and attention based encoder- decoder to recognize real world noisy speech. Long Short Term Memory (LSTM) based language model combined with HMM based speech recognition system built on switchboard and librispeech dataset was exploited by Wei Zhou et al. (Zhou, Schlu¨ter & Ney, 2020) had achieved very low WER. Listen Attend and Spell (LAS) is a combination of recurrent neural network encoder and attention based decoder achieves a WER of 14.1% over Google voice search without an external language model. In Rose S, Kumar L and Renuka D (2019) Speaker independent speech recognition model using visual features were studied. Several research works been carried out by G, A, K, D and Karthika (2020) (MC, Renuka & Kumar, 2021) on speech analytic using different deep learning techniques for various research applications.

3. Methodology

In Audio Visual Speech Recognition (AVSR) there are three modules like audio speech recognition, visual speech recognition and multimodal fusion. The basic architecture of audio visual speech recognition is given in Fig. 2.

3.1. Audio speech recognition (phonemes to text)

The process of converting spoken utterances into text is known as audio speech recognition. Librispeech dataset is used to train the neural network model. The input sound signal is split into sound frames of window size 20- 25 ms long with stride 10 ms. Feature extraction takes audio as input and extract information as features. In general several representation methods are used to convert input speech from one dimensional signal to two dimensional representations. Mel-Frequency Cepstral coefficient (MFCC) and spectrogram are the most frequently used representation for audio signals given in Table 1 and Fig. 3. The logarithm of mel frequency is used to calculate MFCC features by applying discrete cosine transform (DCT).Mel frequency is calculated using equation Eq (1).

$$mel(f) = 2595 * \log\left(1 + \frac{f}{100}\right) \tag{1}$$

Where, mel(f) is the frequency (mels) and f is the frequency (Hz). MFCC is calculated using Eq. (2)

$$c_n = \sum_{n=1}^{k} \log S_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right] \tag{2}$$

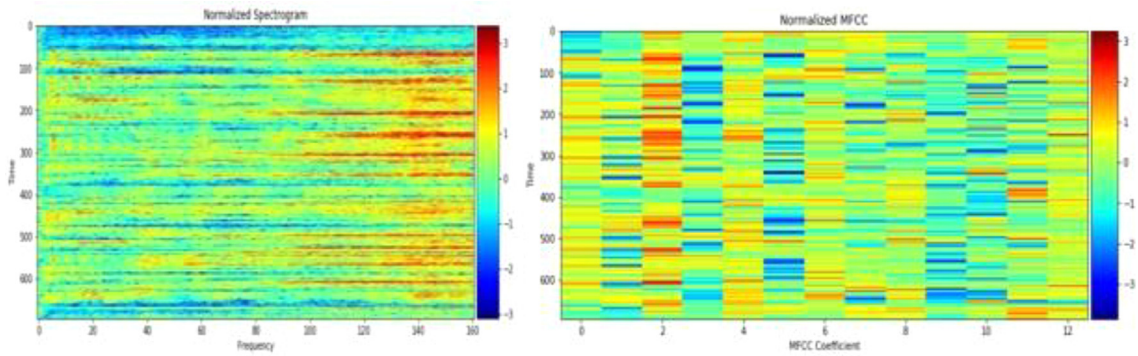


Fig. 3. Spectrogram and MFCC.

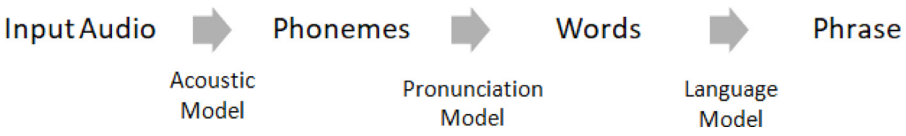


Fig. 4. Components of ASR.

Table 1
Comparison of Spectrogram and MFCC.

Attributes	Spectrogram	MFCC
Frequency	Low and Medium	Low
Filter type	Band pass filter	Mel
Filter Shape	Linear	Triangular
What is Modeled?	Human Auditory System	Human Auditory System
Computation Speed	High	High
Type of Coefficient	Spectral	Cepstral
Noise Resistance	Low	Medium
Sensitivity to Quantization or Additional Noise	Medium	Medium
Reliability	Medium	High

where, k is the number of melcepstrum coefficients, Sk is the output of filter bank and Cn is the final MFCC coefficients.

In this paper we have used spectrogram features to show audio signals in time domain versus frequency domain. Time domain represents audio amplitude over time while frequency domain gives audio intensities over different frequency for each and every frame of the signal. Audio speech recognition consists of three Components such as acoustic model, pronunciation model and language model shown in Fig. 4.

- Acoustic model: It takes speech input and converts to its corresponding phonemes the basic linguistic unit.
- Pronunciation model: Otherwise known as lexicon or dictionary which maps the phonemes to words. For instance the phoneme sequence f-ay-v can be mapped to the word five.
- Language model: To identify the most probable sequence of words or phrase.

The input time sensitive audio data as spectrogram features is down sampled using a single one dimensional convolution layer. In this convolution layer the inputs are multiplied that is convoluted based on kernel size and filter size. The kernel size determines the sliding window size. The idea of convolution is to extract feature for the input audio tensor in a better way. The output from convolution layer is passed into a set of gated recurrent unit which is a variant of recurrent neural network layer followed by a batch normalization layer. Recurrent neural networks are the one which predicts the next possibility using previous output data. Hence it works better for continuous time series data such as speech prediction. Batch normalization is applied between the GRU layers to train the data by performing normalization and to converge quickly. Regularization is introduced by using dropout to extrapolate unseen new data apart from the training data and to avoid overfitting. Clipped ReLu and softmax activation function is used to introduce non linearity in the data. Fig. 5 represents the neural network model workflow of ASR. Table 2 depicts the overall workflow of AVSR system.

Table 2
Pseudo code for AVSR System.

Pseudo code for AVSR System
Input: Audio and Video files
Output: Transcribed text
Module 1: Automatic Speech Recognition
Step 1: Feature Extraction using MFCC and Framing with size 20 ms.
Step 2: Input feature $\{x_1, x_2, \ldots x_n\}$ is fed into stack of GRU layer with softmax activation at end.
Step 3: The output characters $\{y_1, y_2, \ldots y_n\}$ are applied into connectionist temporal classification (CTC) layer with special character blank to eliminate duplicates.
Module 2: Visual Speech Recognition
Step 1: Face detection and Lip localization is performed to detect region of interest
Step 2: Lip images are classified using stack of CNN layers.
Step 3: The identified output characters are then fed into CTC to remove the duplicates.
Step 4: Decision level fusion of audio and visual output from module 1 and module 2 for combining the multimodal results.

Table 3
Prediction Probability Algorithm.

Prediction Probability Algorithm
Input: Images from Grid Dataset
Output: Sampled subset of images Begin
Step1: Create two character model (ab, cd, ef, gh, ij, kl, mn, op, qr, st, uv, wx, yz)
Step2: Select the first image of the dataset
Step3: Calculate the prediction probability of the image using the created two character model
Step4: If the prediction probability is less than the threshold value (0.9) then delete the image else select the image for further processing
Step5: Repeat step 3 and 4 for all the images in the dataset.
End

3.2. Visual speech recognition (visemes to text)

A Visual Speech Recognition system is an automatic process of detecting the spoken words from speaker's lip movement. This technology provides an alternative way of communication (i.e. Visual communication) for people with hearing impaired problem. The dataset can be

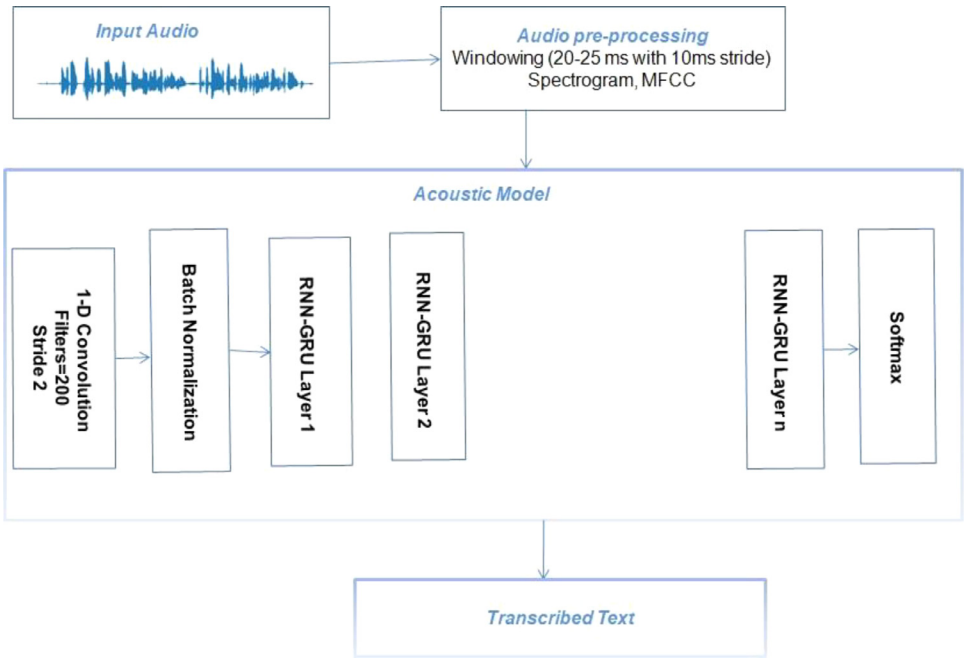


Fig. 5. Neural Network Model of ASR.

generated by using character level or word level. The character level dataset are created with alphabets or phonemes. The word level dataset are created with specific word. It's difficult to train the system with entire words in the particular language. Hence to train the proposed neural network model the concept of character model dataset from Grid corpus is chosen. GRID (Global Research Identifier Database) dataset is used which was collected by Cookee in a recording studio. The various parts in the visual speech recognition system are Face Identification, Lip Localization and Lip Reading System. The face identification is the process of identifying the faces in the image. The face identification can be done by using traditional method or deep learning approaches. The lip localization is the process of extracting the lip or identifying the region of interest (ROI) from the identified face. Two variants of lip localization methods are as follows

- Image based approach: Image based approach is to extract lip region based on the color in a simpler way. Image based approach includes RGB Model, Hue Saturation Value Model and YCbcrcolor model.
- Model based approach: Active Shape Modelling (ASM) and Active Appearance Model (AAM) are the extensively used model based approaches.

In the proposed architecture the videos are divided into frames. The extracted unique frames are given into VSR unit. For instance consider a video speaking "add". The three extracted unique frame have frame of "a", frame of "d" and frame of "d. These frames are given into VSR system to predict the spoken character as shown in Fig. 6.

Our proposed prediction probability algorithm for sampling the dataset is shown in Table 3.

Each VSR unit is developed by two character model totally include 13 VSR units. The lip movement is extracted by using ASM Model. The ASM Model searches for facial landmarks from the mean shape and fits it to the correct landmark position by appropriate movement. The mouth region is represented by 48–68 trait points. The extracted lip is given as input to each VSR Unit. The prediction probability is calculated separately for each VSR unit. The output of VSR unit having highest prediction probability is taken as output of proposed model as shown in Fig. 7. Each VSR unit consists of a stack of CNN and maxpooling layer. The VSR unit is trained with two characters. The spoken character is given to each VSR unit and the probability prediction is calculated as

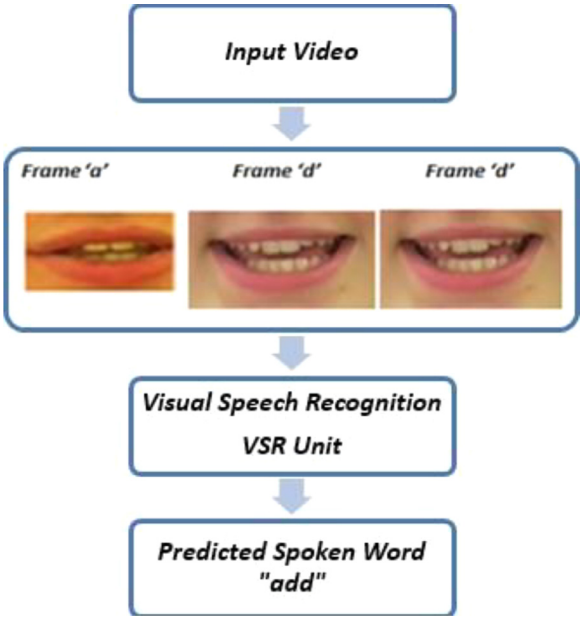


Fig. 6. Architecture of VSR.

shown in Eq. (3).

$$Probability\ prediction\ =\ P\ (X,\ C) \tag{3}$$

where, X is the predicted value and C is the predicted class.

The probability prediction value ranges from 0 to 1. The predicted value is the amount on how much it matches with the classified class. Maximum value among the predicted VSR Unit is selected as output as shown in Eq. (4).

$$Y\ (X)\ =\ max[P\ (VSR1),\ P\ (VSR2),\ P\ (VSR13)] \tag{4}$$

where, k is the number of melcepstrum coefficients, Sk is the output of filter bank and Cn is the final MFCC coefficients.

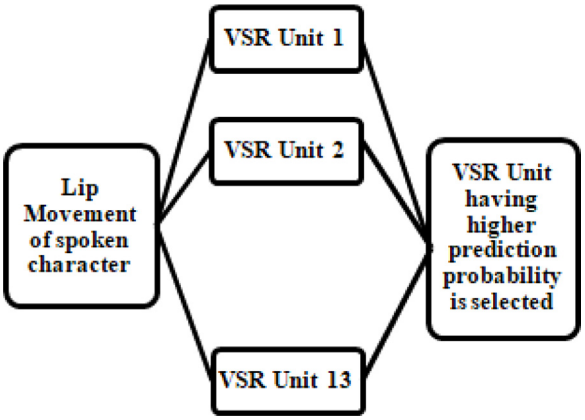


Fig. 7. Proposed Model of VSR.

Table 4
Dataset Description.

Dataset	Description
LibriSpeech	Open source speech dataset. We have used training data with 337 million tokens and testing data with 346 million tokens.
GRID	Open source Audio Video dataset recorded in an acoustic studio.

Table 5
Grid Dataset Description.

Description	Attributes
Command	Bin, lay, Place, set
Color	Blue, Green, Red, White
Preposition	At, by, in, with
Letter	A Z
Digit	0 . . . . . 0.9
Adverb	Again, no, please, soon

Table 6
Comparison of Speech to Text Models.

Models	Layers Used	WER%
Wav2Letter[5]	17 1D-Convolutional Layers and 2 Fully Connected Layers	6.67
DeepSpeech2 (Amodei et al., 2016)	2–3 Convolution layer 3–7 GRU/LSTM layer 1–2 Fully connected layer	6.71
Proposed Model	1 1D Convolution Layer + batch Normalization 5–13 RNN GRU layer	6.59

4. Result analysis

The two dataset used for experimental analysis are LibriSpeech and Grid as shown in table 4 and table 5. And the models are trained on high end workstation featured with GeForce RTX card using tensorflow framework.

In general word error rate is the ratio between number of errors to total number of words. The word error rate (WER) is the evaluation metric used for speech recognition which is calculated using Levenshtein distance as given in Eq. (5).

$$WER = \frac{L(T,P)}{W} = \frac{Sum(Insertion,\ Deletion,\ Substitution)}{W} \quad (5)$$

where, k is the number of melcepstrum coefficients, Sk is the output of filter bank and Cn is the final MFCC coefficients.

RNN-GRU speech to text model is compared against wav2letter+ and Deep- Speech2 architecture as shown in table 6.

The alphabets are extracted from the GRID dataset and ASM model is applied for extracting the lip region. To analyze the results the Lip movement L is given to the all 13 VSR unit i.e. to the stack of CNN and maxpooling layer. The prediction probability is calculated at each VSR unit. Among 13 VSR units the VSR unit 6 has highest prediction probability as shown in Fig. 8. Table 7 shows the proposed lip reading model comparison with other existing state of the art (SOTA) results.

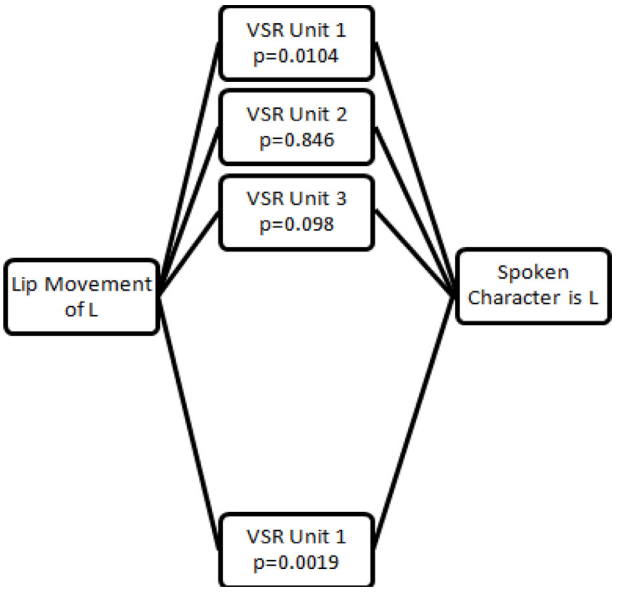


Fig. 8. VSR system for Detecting L.

Table 7
Comparison of Lip Reading Models with SOTA.

Models	WER%
LipNet (Assael et al., 2016)	4.9%
Our Proposed Model Using CNN	4.5%

Conclusion

An enhanced AVSR system will be beneficial to our society to act as an assistive technology for hearing impaired people and also detect speech in noisy environment. The efficient model is built in our proposed framework by using deep learning algorithms. ASM is used for lip localization and CNN based VSR unit is built to boost overall performance. A higher accuracy of about 95% with lesser word error rate of 6.59% is achieved. Combining visual information with speech recognition using deep learning algorithms offers an efficient visual speech recognition system. VSR plays a major role in recognizing speech without audio. It also acts as a complementary tool for Audio Visual Speech Recognition (AVSR) system. In future we are proposing a BERT based language model which can improve the WER of our proposed audio model and also an audio- visual multimodal fusion framework for improving the performance of automatic speech recognition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Department of Science and Technology Interdisciplinary Cyber Physical Systems (DST-ICPS), India.

References

Alothmany, N., Boston, R., Li, C., Shaiman, S., & Durrant, J. (2010). Classification of visemes using visual cues. In *Proceedings ELMAR-2010* (pp. 345–349).

Amodei, Dario, Ananthanarayanan, Sundaram, Anubhai, Rishita, Bai, Jingliang, Battenberg, Eric, et al. (2016). Deep speech 2: End-to-end speech recog- nition in English and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)* (pp. 173–182). JMLR.org.

Assael, Yannis, Shillingford, Brendan, Whiteson, Shimon, & Freitas, Nando. (2016). LipNet: Sentence-level lip-reading.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960–4964).

Collobert, Ronan, Puhrsch, Christian, & Synnaeve, Gabriel. (2016). Wav2Letter: An end-to-end ConvNet-based speech recognition system.

Easton, Randolph, & Basala, Marylu (1982). Perceptual dominance during lipreading. *Perception Psychophysics, 32*, 562–570. 10.3758/BF03204211.

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)* (pp. 681–688). 10.1109/IJCNN.2017.7965918.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of speech and hearing research, 11*(4), 796–804.

Frolov, I., & Sadykhov, R. (2009). Face recognition system using SVM-based clas- sifier. In *2009 IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (pp. 394–399). 10.1109/IDAACS.2009.5342952.

Pooventhiran, G., Sandeep, A., Manthiravalli, K., Harish, D., & Renuka, Karthika (2020). Speaker-independent speech recognition using visual features. *International Journal of Advanced Computer Science and Appli- cations, 11*. 10.14569/IJACSA.2020.0111175.

Gordan, M., Kotropoulos, C., & Pitas, I. (2002). A support vector machine based dynamic network for visual speech recognition applications. *EURASIP Journal on Advances in Signal Processing, 2002*, Article 427615. 10.1155/S1110865702207039.

Heckmann, Martin Kroschel, Savariaux, Kristian, Berthommier, Christophe, & Fred, Eric (2002). DCT-based video features for audio-visual speech recognition. *7th International Conference on Spoken Language Processing.*

Lopes, Carla, & Perdig˜ao, Fernando. (2011). Phone recognition on TIMIT database. doi: 10.5772/17600

Shunmugapriya, M. C., Renuka, Dr. D. Karthika, Kumar, Dr. L. Ashok, et al. (2021). Recurrent network-based hybrid acoustic model for automatic speech recognition. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(10), 7308–7315. 10.17762/turcomat.v12i10.5621.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). 10.1109/ICASSP.2015.7178964.

Pandey, Laxmi, & Arif, Ahmed Sabbir (2021). LipType: A silent speech recognizer augmented with an independent repair model. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing System: 1* (pp. 1–19). New York, NY: Association for Computing Machinery. Article. 10.1145/3411764.3445565.

Petajan, E., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988). An improved automatic lipreading system to enhance speech recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88* (pp. 19–25). New York, NY: Association for Computing Machinery. 10.1145/57167.57170.

Puviarasan, N., & Palanivel, S. (2011). Lip reading of hearing impaired persons using HMM. *Expert Syst. Appl. 38, 4*(April 2011), 4477–4481. 10.1016/j.eswa.2010.09.119.

Rose, Lovelyn S, Kumar L, Ashok, & Renuka D, Karthika (2019). *Deep learning using python.* India: Wiley.

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)* (pp. 19–25). 10.1109/IWDRL.2018.8358210.

Torfi, A., Iranmanesh, S. M., Nasrabadi, N. M., & Dawson, J. M. (2017). 3D convolutional neural networks for cross audio-visual matching recognition. *IEEE Access : Practical Innovations, Open Solutions, 5*, 22081–22091.

Vakhshiteh, Fatemeh, Almasganj, Farshad, & Nickabadi, Ahmad. (2018). Lip- reading via deep neural networks using hybrid visual features. *Image Analysis Stereology, 37*, 159. 10.5566/ias.1859.

Wang, S. H., Wu, K., Chu, T., Fernandes, S. L., Zhou, Q., Zhang, Y. D., et al. (2021). SOSPCNN: Structurally optimized stochastic pooling convolutional neural network for tetralogy of fallot recognition. *Wireless Communications and Mobile Computing*, 2021.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017, December). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1240–1253. 10.1109/JSTSP.2017.2763455.

Zeghidour, N., Xu, Q., Liptchinsky, V., Usunier, N., Synnaeve, G., & Collobert, R. (2018). Fully convolutional speech recognition. *ArXiv* abs/1812.06864.

Zhang, Yu-Dong, Satapathy, Suresh, Guttery, David, Gorriz, Juan, & Wang, Shuihua (2021). Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Information Processing and Management, 58*, Article 102439. 10.1016/j.ipm.2020.102439.

Zhou, W., Schlu˜ter, R., & Ney, H. (2020). Full-sum decoding for hybrid hmm based speech recognition using LSTM language model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7834–7838).