GAP ANALYSIS AND DEVELOPMENT OF AN AUTOMATIC SPEECH RECOGNITION FRAMEWORK FOR HEARING IMPAIRED PERSON BY USING HYBRID TECHNIQUES

Prof (Dr) Sanjay Kumar[1],
Department of Computer Application
Shri Ram Murti Smarak College of Engineering &Technology, Bareilly-UP

Prof (Dr) Manoj Pandey[2]
SRMS International Business School, Unnao-UP

Prof (Dr) Devendra Kumar Pandey[3]
Faculty of Management Studies
Medi-Caps University, Indore

Dr Anand Kumar Srivastava[4], Associate Professor
Amity University Madhya Pradesh, Gwalior

Harshit Kumar[5], B.Tech-CSE. Student
Department of Computer Science & Engineering
Madan Mohan Malaviya University of Technology. Gorakhpur.-UP
Corresponding author: Sanjaysatyam786@gmail.com[1]

ABSTRACT

People with hearing loss would greatly benefit from assistive technology if it used Audio Visual Speech Recognition (AVSR). 466 million individuals worldwide are deaf or partially deaf. They use lip reading to grasp what is being said since they are deaf or hearing impaired. Many students with hearing loss struggle because of a scarcity of qualified sign language facilitators and the hefty price tag on assistive technology. Using cutting-edge deep learning models, we've developed a new way for visual voice detection. In addition, the present VSR approaches have flaws that need to be corrected. Our new method merges audio and visual speech outputs as a result. An audio-visual speech recognition model that incorporates deep learning has been proposed to improve lip reading speed and accuracy. According to this study, the system's performance has been significantly improved, with word error rates of 6.59 % for ASR and 95 % for lip reading.

Key words: Acoustic, Speech, voice, phonemes, learning techniques, NLP

INTRODUCTION

Artificial Intelligence (AI) and natural language processing (NLP) are utilized by the speech recognition system to recognize words. Those who are deaf or hard of hearing must rely on lip reading as their primary way of communicating. The current state of affairs for children who are deaf or hard of hearing is shown in Figure 1. To be successful in school, many deaf and hard-of-hearing kids rely on the use of sign language. The obstacles are discussed from both the facilitator's and the student's perspectives. As a result of this research, we have developed a lip-reading system that is more accurate than any other currently available assistive technology.

It is an automated technique of turning audio characteristics into text. The most often used datasets for automated voice recognition are Librispeech and Timit [1]. These deep learning-based ASR architectures include Deep Speech, LAS, and Wav2Letter [5], which all have better recognition performance than other ASRs. Because it does not need an audio context, visual speech recognition has become more important in speech recognition systems in recent years. Automated lip movement tracking is used to recognize spoken words in a Visual Speech Recognition system which allows those with hearing problems to communicate with others using this technology (i.e., visual communication). The audio speech recognition system is greatly enhanced by VSR, a crucial component of the Auditory Visual Speech Recognition System (AVSR). In today's world, VSR devices are often used in scenarios like as driving a vehicle or talking on a cell phone in the open air. Traditional statistics and machine learning methods coexist with a deep learning approach in the VSR system. The conventional technique has a high proportion of spelling and grammar mistakes. Deep learning has been used by researchers to create a novel method for decreasing the number of word errors. Deep learning VSR architecture[6] such as lip net and large-scale visual speech recognition are popular. Traditional techniques have a lower word mistake rate and lower recognition accuracy than deep learning systems.

The following is the list of the paper's key contributions:

- Using survey data from various hospitals, clinics and other primary sources, it has been observed the gaps are there in present VSR technique on account of maintenance cost, dubious products, complex working system and age factors.
- Using Recurrent neural networks (RNN)-GRU and convolution neural networks (CNNs), we've developed a voice to text model that can convert spoken words into text.
- An audio-visual speech-recognition fusion has been developed in contrast to the standard system.

RELATED WORK

By tracking the movement of the speaker's lips, lip readers may understand what is being said. Lip reading is an automated method in Visual Speech Recognition. Lip reading methods rely on Visages as their primary visual unit, whereas phonemes serve as the foundational unit of language. But people with hearing loss have a hard time recognizing spoken words over lengthy periods of time [7]. An individual's height ratio is used for the first VSR technique [2] and as a result, scientists are concentrating their efforts on a technology known as visual speech recognition, which automatically detects when someone is speaking by looking at their lips [15]. There was an earlier version that had some of these character traits: mutual knowledge, quality features, and the appearance of the tongue and teeth. Lip images can be categorized using additional machine learning classifiers, such as support vector machines and hidden Markov models [18,9] As an example, convolution neural networks (CNNs) are widely used in pattern recognition and medical applications because of their cutting-edge deep learning models [17]. To improve identification accuracy and minimize the number of word mistakes, deep learning methods have recently been emphasized in speech analytic applications [3] produced an improved version of lipnet called the Deep Learning Architecture lip net, which Oxford University developed in 2017 together with liptype [13].

Variable-length audio signals are translated by audio speech recognition (ASR) systems into a variable-length word sequence. No matter how well-known voice recognition algorithms like HMM[4] and GMM are used, there appears to be a significant failure rate for syllables. A sequence of states is used in this statistical approach to deal with audio inputs that change in duration and temporal structure. It was shown that the word error margin of a convolutional language model trained on the Wall Street Journal and librispeech datasets was greater when the noise was high. [7], CTC and an attention-based encoder-decoder hybrid model were created to detect chaotic speech in the actual world by [10]. Other researchers, such as Wei Zhou et al., have used HMM-based voice recognition based on datasets from the switchboard and librispeech to obtain very low WER. Google voice search has a WER of 14.1 percent without using an external language model, owing to LAS, a recurrent neural network encoder, and an attention-based decoder. [19] created a speech recognition model that is speaker-independent and uses visual information. G.A., K.D., and Karthika used several deep learning algorithms to undertake a series of research projects on speech analysis [14].

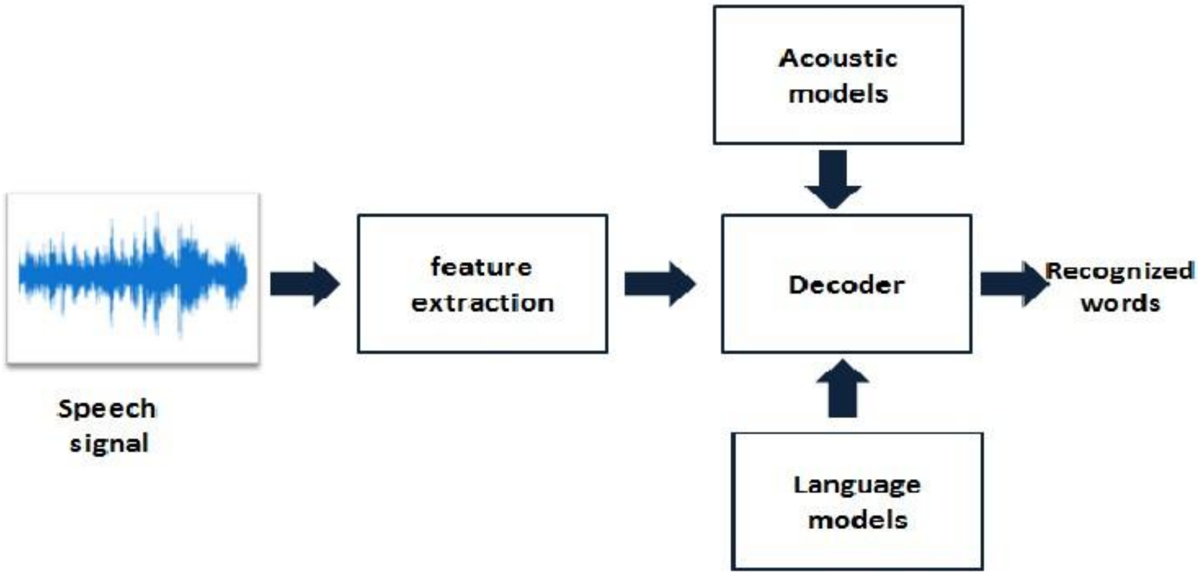


Fig 1:Architecture of ASR

METHODOLOGY

The three main components of Audio Visual Speech Recognition are multimodal fusion[4], audio speech recognition, and visual speech recognition. Figure 1 depicts an auditory visual representation of speech recognition.

Audio speech recognition

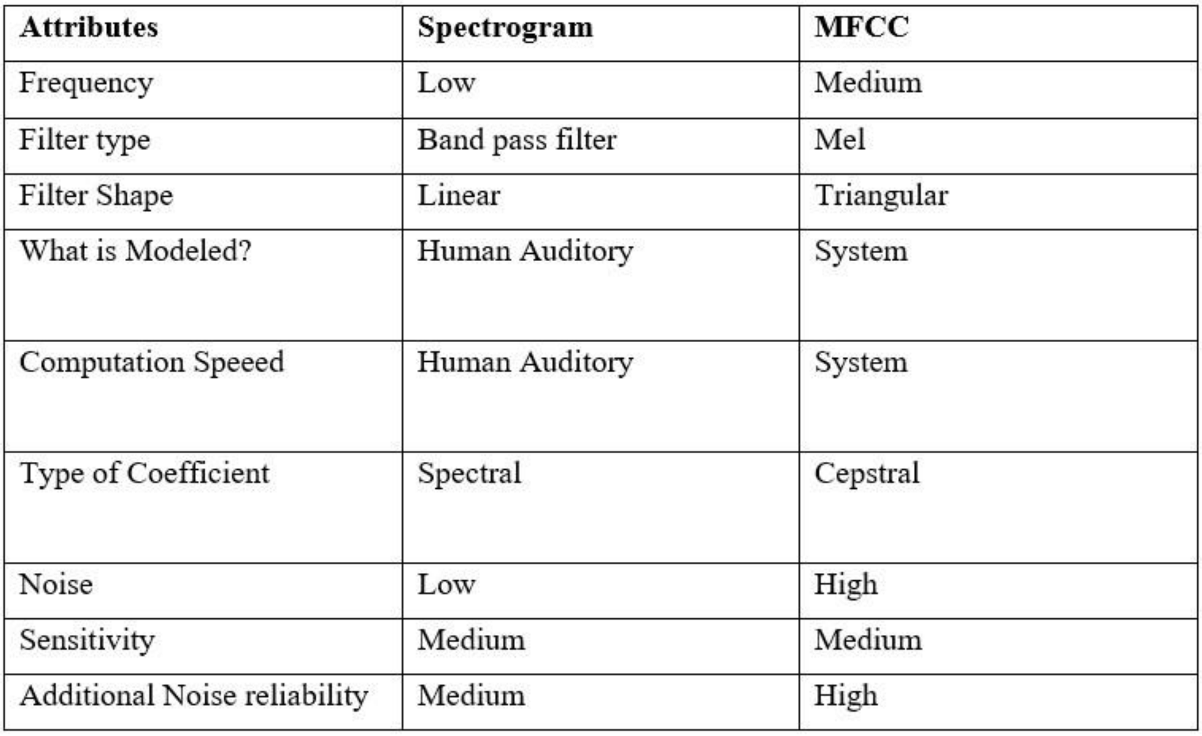
The words spoken by a person are translated into text using audio voice recognition software. The Librispeech data set is used to develop the ‘Neural Network Model’. Windows of 20-25ms and a stride of 10ms are used to divide the input signal into sound frames. Audio is sent into the feature extraction process, which then separates out the various features. There are a variety of representation approaches available for transforming one-dimensional signals into two dimensional images from spoken input. Mel-Frequency[11] Most often, audio signals are represented in Table 1 and Fig. 3 by Cepstral coefficient (MFCC) and spectrum, respectively. To determine MFCC features, the logarithm of the mel frequency is used to discrete cosine transform (DCT). Using the following equation , the Mel frequency is determined (1):

$mel(f)$ = 2595$*$ log (1 + f100)……….. (1)

mel(f) is a measure of frequency (measured in millihertz; mels);

f is frequency (Hz). Eq. 2 is used to calculate MFCC

cn = ∑k n=1log Sk cos[n(k − 12) πk] …………(2)



Attributes	Spectrogram	MFCC
Frequency	Low	Medium
Filter type	Band pass filter	Mel
Filter Shape	Linear	Triangular
What is Modeled?	Human Auditory	System
Computation Speeed	Human Auditory	System
Type of Coefficient	Spectral	Cepstral
Noise	Low	High
Sensitivity	Medium	Medium
Additional Noise reliability	Medium	High

Table 1: Comparison of Spectrogram and MFCC

The number of melcepstrum coefficients is k, the filter bank output is Sk, and the number of melcepstrum coefficients is k. k=cn

The length and sensitivity of audio samples[12] were compared using spectrogram characteristics in this research. The amplitude of an audio signal is expressed in terms of time, whereas the frequency of the signal is expressed in terms of frequency.

The acoustic model, the pronunciation model, and the language model are all components of audio speech recognition (ASR), as illustrated in Fig. 1.

- **Acoustic model:** A phoneme is a fundamental linguistic unit that is generated from the input of speech.
- **Pronunciation model:** A phonetic lexicon or dictionary is another term for this tool. This may be done by mapping phoneme patterns to words like "five," for example.
- **Language model:** Determine the sequence a group of words or a phrase is most likely to occur in.

One-dimensional convolution is used to extract spectrogram characteristics from the time sensitive audio data that is supplied into it. As the kernel and filter sizes change, this layer of convolution multiplies and combinational circuits the inputs accordingly. The size of the sliding window is determined by how large the kernel is. As an audio processing technique, convolution may be used to increase extracting features quality. It is necessary to perform batch normalization[16] once the convolution layer's output is processed through a collection of gated recurrent units (a variation of the recurrent neural network layer). Recurrent neural networks use prior output data to make predictions about the future.

Prediction probability algorithm	**Results**
Input	Images from Grid Dataset
Output	Sampled subset of images begin
Step 1	Create two-character model (ab, cd)
Step 2	Select first image of the dataset
Step 3	Calculate the prediction probability image using created character model
Step 4	If the prediction probability is less than the threshold 0.9 then further process of image takes place
Step 5	Step 3 and 4 repeated for all the images in dataset End

Table 2: Probability Prediction Algorithm

As a consequence, this strategy is effective for dealing with time series data, such as speech prediction. By normalizing the input, it may be utilized to train the model and quickly converge between the GRU layers.

Dropout is used to extrapolate new data not found in the training data and prevent overfitting, which allows normalization to be introduced. Non linearity is introduced into the data using clipped ReLu and softmax activation functions. In Fig.2, ASR's neural network model process[20] is shown.

Visual speech recognition

A person's lips move in a way that conveys what they are saying while they are communicating. Deaf or hard-of-of-hearing people now have an alternative method of communication thanks to this new technology (i.e., visual communication). Character- or word-level data can be utilized to generate the dataset. Using alphabets or phonemes, the datasets at the character level are compiled. The datasets at the word level are built using a predefined collection of words. System training using whole words in a certain language is difficult. As a result, the neural network model supplied here is trained using the character model dataset from Grid corpus[18]. As part of this study, the data from the Cookee recording studio is utilized, as well as the data from the Global Research Identifier Database (GRID). This system is comprised of different components, including Face Detection, Lip Localization and Lip Reading. The technique of recognising the faces in an image is known as face identification. Both classic and advanced methods may be used to identify people's faces. As the name suggests, lip localization involves removing the lip or determining its location on a face. The following is a list of two liplocalization methods.

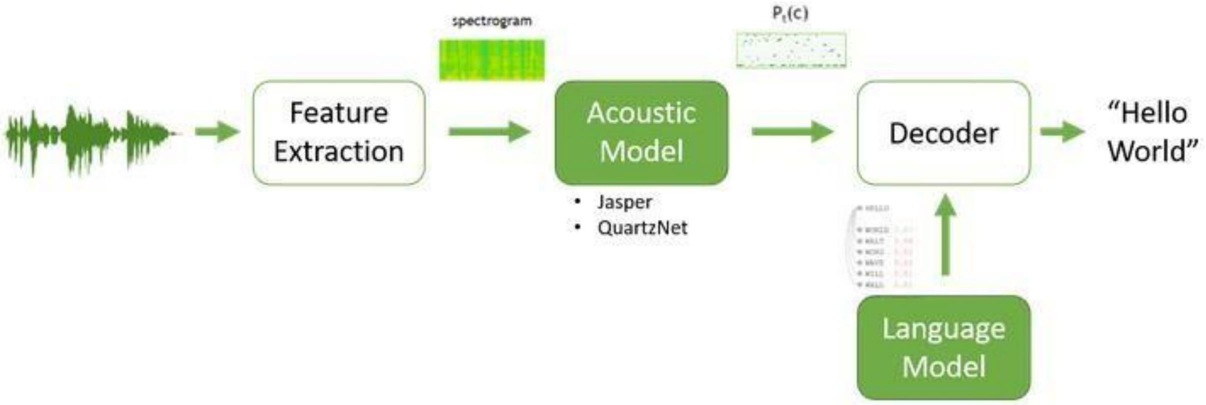


Fig2: ASR model

- **Image based approach:** A more simple method of identifying the lip area is to use an image based technique that relies on the colour of the image. An image-based method may be found in the YCbcrcolor model, the RGB model, and the Hue Saturation Value model.
- **Model based approach:**'Active shape modelling' (ASM) and 'active appearance modelling' (AAM) are two model-based techniques that are often used. Frames are used to separate the video clips in the proposed layout. The unique frames are sent to the VSR unit. A video of someone saying "add" is an example. Frames of "a,"“a,” and "d" are among the three retrieved unique frames. In order to forecast the spoken character, the VSR system takes these frames as input. Fig. 3 depicts our suggested sampling algorithm for the dataset.

There is a total of 13 VSR units produced by two-character models for each VSR type. ASM Model is used to extract the lip movement. The ASM Model uses the mean shape to find face landmarks and then moves the mean shape to match the right landmark location. A total of 48– 68 trait points are assigned to the area around the mouth. Each VSR Unit receives the extracted lip as an input. It is possible to predict the outcome for each VSR independently by calculating the prediction probability. According to Fig.2, the predicted output of the proposed model is derived from the output of the VSR unit with the greatest prediction probability. A CNN and maxpooling layer stack forms the basis of each VSR unit. Two characters are used in the training of the VSR. Every unit of VSR has its own prediction probability calculated.

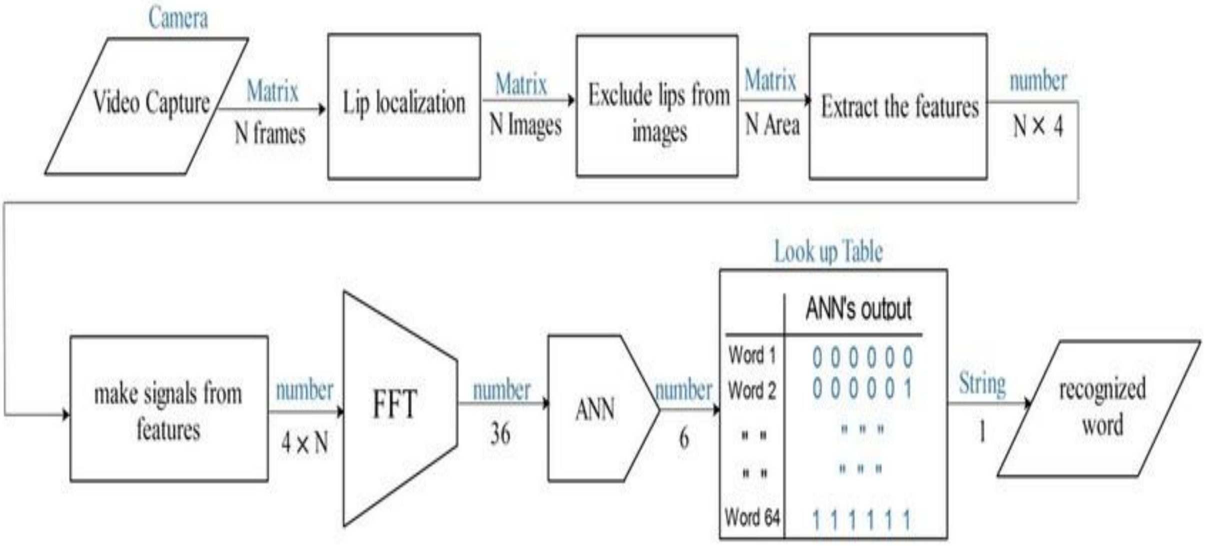


Fig3: Architecture of VSR

VSR's architecture is shown in Equation (3).

$Probabilityprediction = P\ (X, C)$ (3)

For instance, X represents the expected value and C represents the predicted class. 0 to 1 is the range of the probability prediction value. Based on how well it matches its class, a projected value is computed. As indicated in Equation, the VSR unit with the highest anticipated value is selected as the output (4).

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

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

DATASET	DESCRIPTION
Librispeech	We have used training data with 337 million tokens and testing data with 346 million tokens through an Open source speech dataset.
GRID	In an acoustic studio, an open source Audio Video dataset is recorded.

Table 3: Dataset Description

DESCRIPTION	RESULTS
Command	Bin, lay, set, place
Color	Blue, Green, Red, white
Preposition	At, in, with

Letter	A – Z
Digit	0-0.9
Adverb	No, again, please

Table 4: Gridset Description

Tables 3 and 4 illustrate experimental datasets, LibriSpeech and Grid, respectively. Using the tensor flow framework, the models are trained on top-of-the-line workstations with GeForce RTX cards.

The error rate is calculated by dividing the number of words in a text by the total number of words. Levenshtein distance (WER) is used to calculate speech recognition's word error rate (WER).

$WER = L(T , P)$ ………..(5)

$W = Sum(Insertion, Deletion, Substitution)\ W$

Filter output is Sk, which is the final MFCC output with the number k melcepstrum coefficients, followed by Cn.

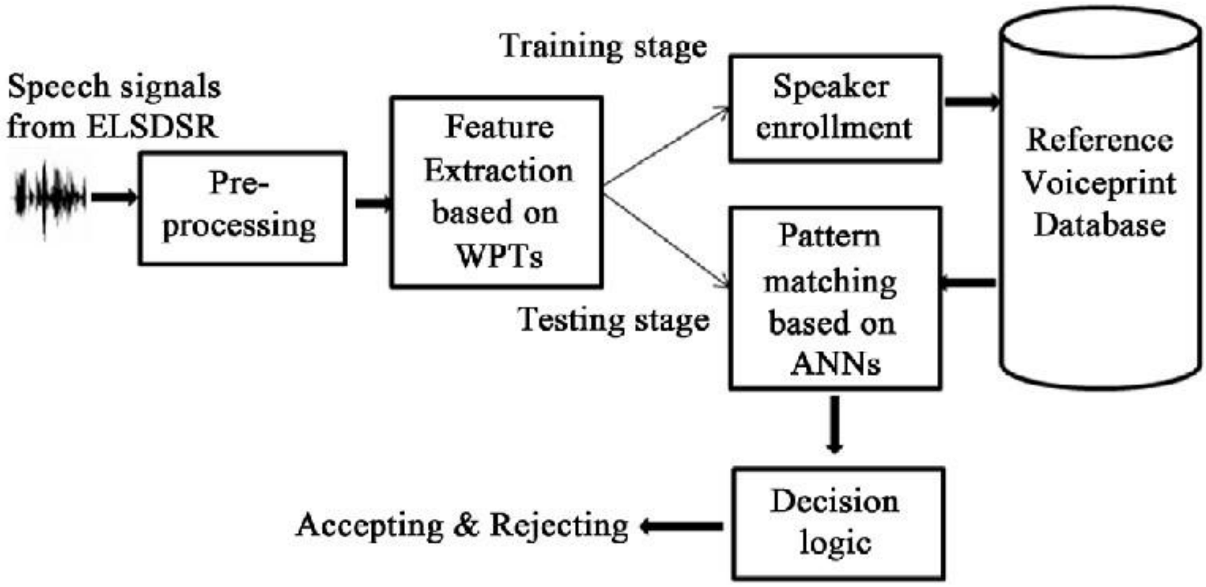

Training stage
Speech signals from ELSDSR
Pre-processing
Feature Extraction based on WPTs
Speaker enrollment
Reference Voiceprint Database
Pattern matching based on ANNs
Testing stage
Decision logic
Accepting & Rejecting

Fig 4: Proposed model of VSR

RESULT ANALYSIS

Tables 3 and 4 illustrate experimental datasets, LibriSpeech and Grid, respectively. Using the tensor flow framework, the models are trained on top-of-the-line workstations with GeForce RTX cards.

The error rate is determined by dividing the total number of words in a text by the number of words in the text. Levenshtein distance (WER) is used to calculate speech recognition's word error rate (WER).

$WER = L(T\ , P)$ ……….(5)

$W = Sum(Insertion, Deletion, Substitution) \quad W$

Filter output is Sk, which is the final MFCC output with the number k melcepstrum coefficients, followed by Cn.

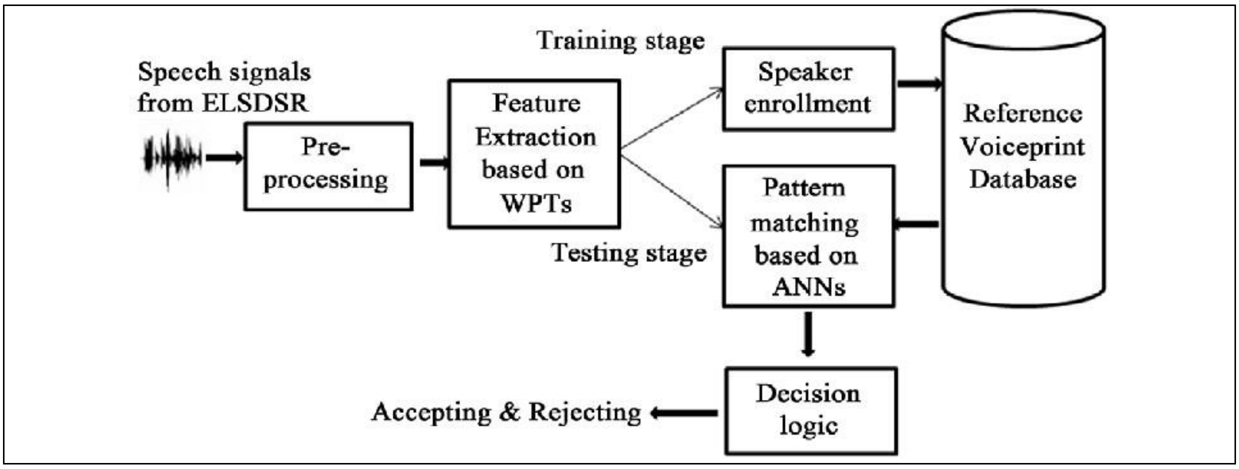

Speech signals from ELSDSR
Pre-processing
Feature Extraction based on WPTs
Training stage
Testing stage
Speaker enrollment
Pattern matching based on ANNs
Reference Voiceprint Database
Decision logic
Accepting & Rejecting

Fig 4: Proposed model of VSR

MODELS	**WER** %
LipNet	4.9%
Our Proposed Model using CNN	4.5%

Table 5: Comparison of the Lip Reading Model with SOTA

The entire stack of CNN and maxpooling layers is provided by movement L. Prediction probabilities are calculated on a per-VSR basis. VSR unit 6 has the greatest prediction probability among the 13 VSR units, as illustrated in Figure. Table 5 compares the findings of our proposed lip reading model to those of previous research.

CONCLUSION

Our society will benefit from assistive technology for the deaf and from enhanced speech recognition in noisy conditions if AVSR is further developed. Using deep learning methods, we have developed a model that is both effective and efficient. Lip localization is handled by ASM, and overall performance is improved by a CNN-based VSR unit.

More than 95% accuracy is attained with just a 6.59 percent word error rate. Deep learning methods for visual voice recognition combine visual information with speech recognition to provide an effective visual speech recognition system. VSR has a significant impact on speech recognition when audio is not available.

To be used in conjunction with Audio Visual Speech Recognition (AVSR). An audio-visual multimodal interaction framework and a BERT language model for enhanced automatic speech recognition performance are both currently under development.

REFERENCES

1. Alothmany, N., Boston, R., Li, C., Shaiman, S., &Durrant, J. (2010). Classification of visages using visual cues. In Proceedings ELMAR-2010 (pp. 345–349).
2. Amodei, Dario, Ananthanarayanan, Sundaram, Anubhai, Rishita, Bai, Jingliang, Battenberg, Eric, et al. (2016). Deep Speech 2: End-to-end speech recognition 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
3. Assael, Yannis, Shillingford, Brendan, Whiteson, Shimon, & Freitas, Nando. (2016). LipNet: Sentence-level lip-reading.
4. 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).
5. Collobert, Ronan, Puhrsch, Christian, & Synnaeve, Gabriel. (2016). Wav2Letter: An end to end ConvNet-based speech recognition system.
6. Feng, W., Guan, N., Li, Y., Zhang, X., & Luo, Z. (2017). Audiovisual speech recognition with multimodal recurrent neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 681–688). 10.1109/IJCNN.2017.7965918.
7. 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.
8. 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 Systems: 1 (pp. 1–19). New York, NY: Association for Computing Machinery. Article. 10.1145/3411764.3445565
9. 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 Applications, 11. 10.14569/IJACSA.2020.0111175.

10. 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.
11. Rose, Lovelyn S, Kumar L, Ashok, &Renuka D, Karthika (2019). Deep learning using Python India: Wiley.
12. 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.
13. 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 Representational Learning (IWDRL) (pp. 19–25). 10.1109/IWDRL.2018.8358210.
14. 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.
15. 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.
16. 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.
17. 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.
18. Zeghidour, N., Xu, Q., Liptchinsky, V., Usunier, N., Synnaeve, G., & Collobert, R. (2018). Fully convolutional speech recognition. ArXiv abs/1812.06864.
19. Zhang, Yu-Dong, Satapathy, Suresh, Guttery, David, Gorriz, Juan, & Wang, Shuihua (2021). Improved breast cancer classification through combining graph convolutional networks and convolutional neural network. Information Processing and Management, 58, Article 102439. 10.1016/j.ipm.2020.102439
20. 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).