Analysis of Audio Visual Feature Extraction Techniques for AVSR System

Dr.L.Ashok Kumar[1], Dr.D.Karthika Renuka[2], M.C.Shunmuga Priya[3]
{lak.eee@psgtech.ac.in[1] ,dkr.it@psgtech.ac.in[2], shunmugapriya.mc@gmail.com[3]}

Professor, Department of Electrical and Electronics Engineering, PSG College of Technology[1],Associate Professor, Department of Information Technology, PSG College of Technology[2],3Research Scholar, Information Technology, PSG College of Technology[3]

Abstract. Recently, Speech technology has started to change the way the human communicate with devices. Speech as a medium of man-machine interaction has been gaining its importance in the modern computer/mobile era. As a result, many computer programs incorporate cutting-edge voice technologies to do diverse jobs. Feature extraction is the foremost step in speech recognition. Understanding the importance of feature extraction methods, this work demystifies the concept behind feature extraction techniques used in audio visual speech recognition.

Keywords: Audio Visual Speech Recognition, Lip Reading Features, Acoustic Features, Feature Extraction.

1 Introduction

In general, Automatic Speech Recognition (ASR) system consist of four modules; Feature Extraction, Acoustic Model, Language Model and Pronunciation Model. Many research ideas have been proposed on these ASR models. In this paper, we made a detailed survey on various audio and visual feature extraction techniques involved in automatic speech recognition.

In general, ASR system the input speech signal is first given into feature extraction module. Speech is a 1D signal which is converted into 2D signal and chopped into 20-25ms sized frames with 10ms shift. In ASR, feature extraction is a process of converting those speech frames into feature vectors. Those feature vectors are fed into the acoustic model which maps the acoustic feature to corresponding basic linguistic unit called phonemes. The phonemes are further improved into words and sentences using pronunciation and language models. Figure 1 shows the pipeline for automatic speech recognition system.

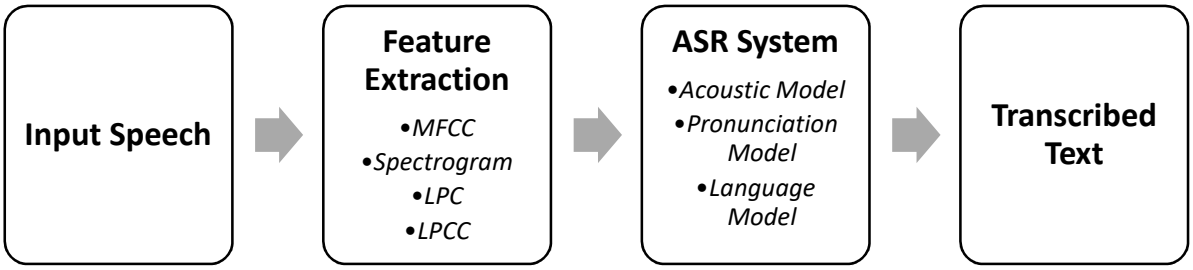

Figure 1: Audio Speech Recognition

Several works on ASR system were carried out in the past decade using machine learning and deep learning models[1]. Deep speech and Listen Attend Spell is a end to end ASR models using neural network techniques [2][3]. It poses several challenges such as ambient noise, speaking style. To overcome the issues involved in ASR system, Visual Speech recognition (VSR) involves the process of tracking lip movement into text [4][5][6][7]. Visemes are the basic visual unit for VSR. Several VSR systems were developed like LipNet, LipType using neural network models[8][9][10].This technology provides an alternative way of communication (i.e. Visual communication) for people with hearing impaired problem.

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314528

2 Feature Extraction Techniques

Figure 2 depicts the variants of feature extraction techniques involved in audio visual speech recognition process.

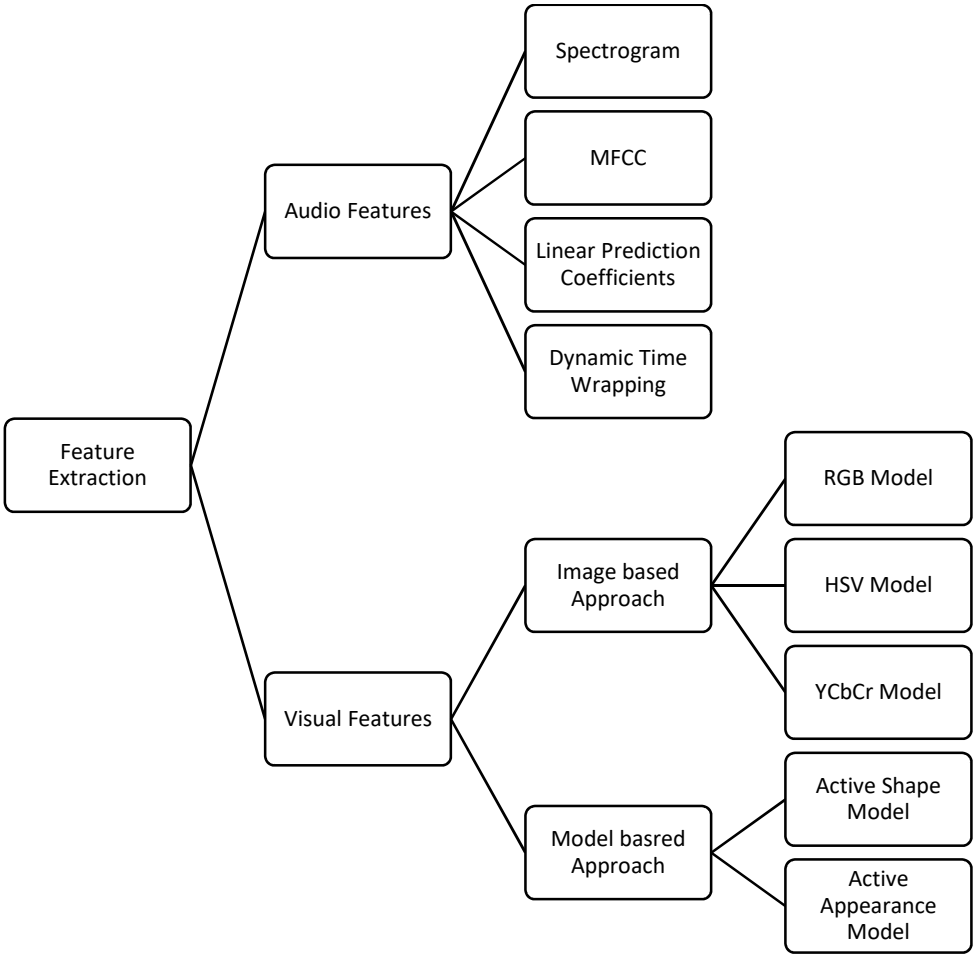

Figure 2: AVSR Feature Extraction Techniques

2.1 Audio Feature Extraction

2.1.1 Spectrogram

Spectrogram is a two dimensional representation of input image. Speech signals are plotted using time in x axis and frequency in y axis as shown in figure 3.

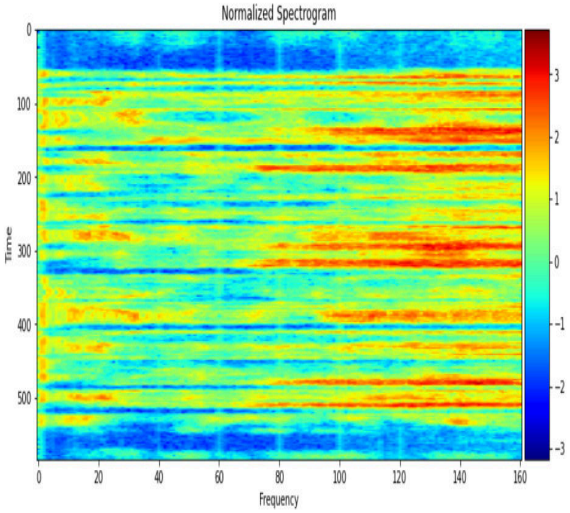

Figure 3: Spectrogram

2.1.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is the benchmark technique for extracting speech features. It replicates the human ear working mechanism. Initially, speech signal is chopped into frames and then power spectrum is obtained by Fast Fourier Transform (FFT) on each frame. The workflow of MFCC coefficient extraction is depicted in figure 4. The mels for any frequency is calculated using the equation 1.

$$mel(f) = 2595\, x\, \log_{10}(1 + \frac{f}{700}) \qquad (1)$$

The MFCC coefficients are calculated using the equation 2.

$$\hat{C}_n = \sum_{n=1}^{k} (\log \hat{S}_k) \cos[n(k - \frac{1}{2}) \frac{\pi}{k} \qquad (2)$$

Figure 5 is the sample MFCC representation. For the speech signals with background noise MFCC feature does not suit well for robust ASR system.

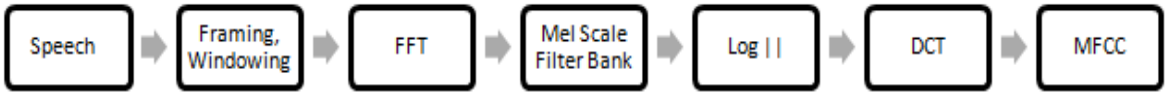

Figure 4: MFCC

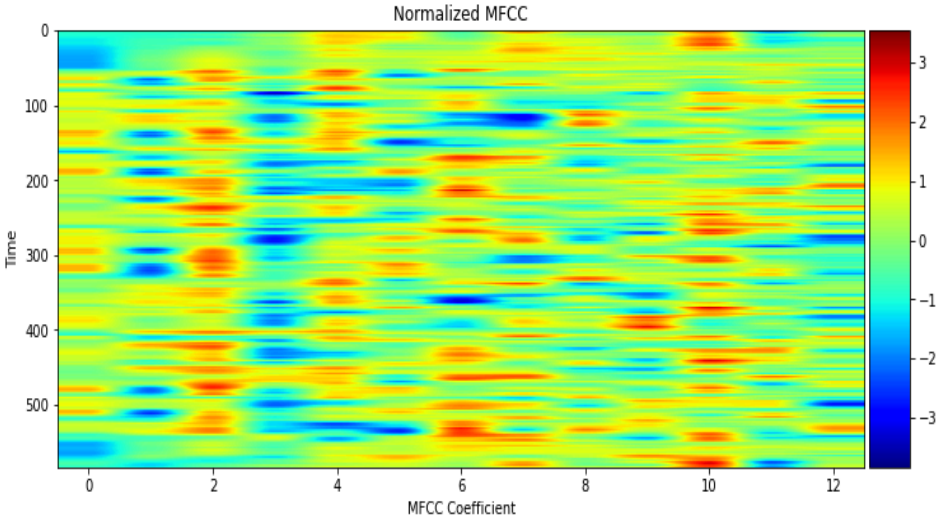

Figure 5: Sample MFCC

2.1.3 Linear Prediction Coefficients (LPC)

LPC feature extraction method is inspired by the human speech production system and yield features of input acoustic signal. Figure 6 is the pipeline of LPC technique. In this technique, speech samples are estimated to be linear combinations of samples from the preceding speech. It is a frame based analysis of speech signals which in turn provides the observation vectors of speech. The input speech signal is digitized and spectrally flattened. The next step is frame blocking where the acoustic signal is chopped into N frames. Then, frames are fed into hamming window to remove the signal discontinuities followed by an auto correlation. Durbin's method is widely accepted where each frame of autocorrelations is converted into an LPC parameter set. The LPC of speech input is given in equation 3. Figure 7 gives the sample LPC plot.

$$\hat{s}\,(n) = \sum_{k=1}^{p} a_k\, s\,(n - k) \qquad (3)$$

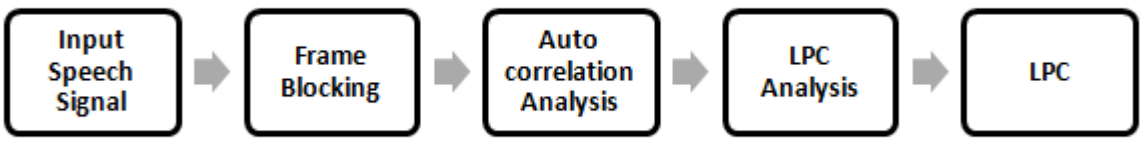

Figure 6: LPC

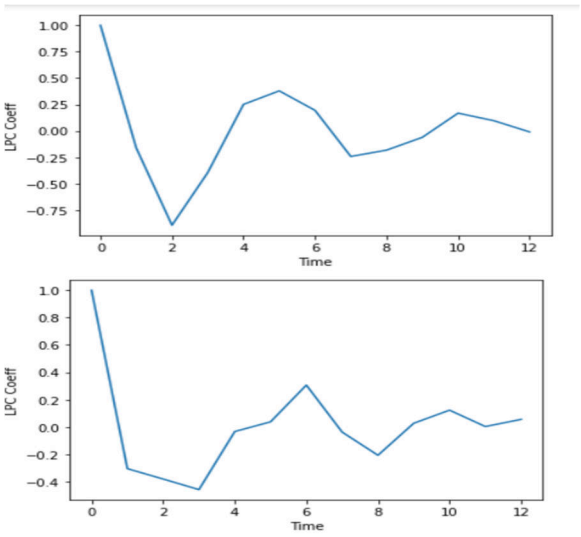

Figure 7: Sample LPC

2.1.4 Linear Prediction Cepstral Coefficients (LPCC)

LPC is employed to calculate the spectrum of the signal. LPCC features yield robust coefficients when compared to MFCC when the input speech is noisy. Steps involved in LPCC are shown in figure 8.

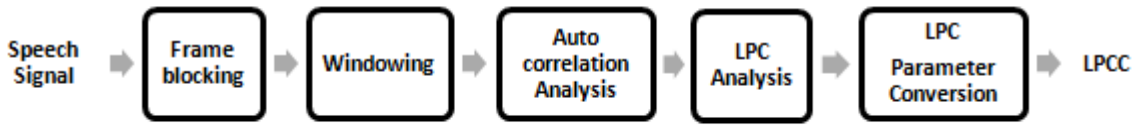

Figure 8: LPCC Processor

LPCC is calculated using equation 4.

$$C_m = a_m + \sum_{k=1}^{m-1} \left[\frac{k}{m}\right] c_k \, a_{m-k} \quad (4)$$

where a_m is the linear prediction coefficient, C_m is the cepstral coefficient. When compared to MFCC, LPCC features are less prone to noise so that they yield lesser word error rate.

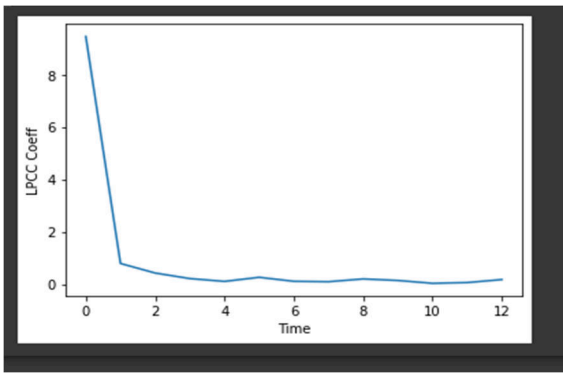

Figure 9: Sample LPCC

2.1.5 Discrete Wavelet Transform (DWT)

The DTW is based on dynamic programming algorithm and extensively used technique in speech. In the time and frequency domains, DWT is a commonly utilised signal processing technique. DWT, is a sub band signal breakdown extension of WT. Equations 5 and 6 show the DWT's scaling and wavelet functions

$$\varphi(t) = \sum_{n=0}^{N-1} h[n]\sqrt{2}\varphi(2t - n) \quad (5)$$

$\rho(t) = \sum_{n=0}^{N-1} g[n]\sqrt{2}\varphi(2t - n)$ (6)

THE DWT for continuous signal is given in equation 7

$(DWT)(m,p) = \int_{-\infty}^{+\infty} x(t).\varphi_{m,p}dt$ (7)

In equation 8, $\varphi_{m,p}$ is the wavelet function

$\varphi_{m,p} = \frac{1}{\sqrt{a_0^m}}\varphi\left(\frac{t - pb_0a_0^m}{a_0^m}\right)$(8)

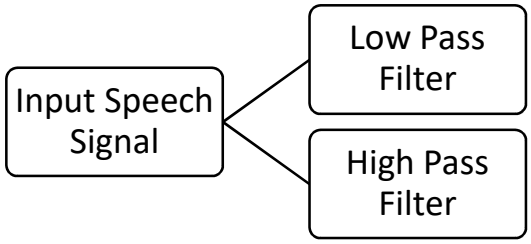

Figure 10: DWT

2.1.6 Perceptual Linear Prediction (PLP)

PLP arose from the non-linear bark scale. The figure 11 shows the steps involved to identify the PLP feature. The speech is first windowed, and then the FFT and square of magnitude are computed, yielding power speech estimations. The bark scale filter bank is then used to combine the power spectrum's overlapping critical band filter response. On the bark scale, this spectrum frequency domain convolution allows low frequencies to conceal high frequencies. The autocorrelation coefficients are obtained using an inverse Discrete Fourier transform. Finally, the auto aggressive coefficients are obtained using spectral smoothing.

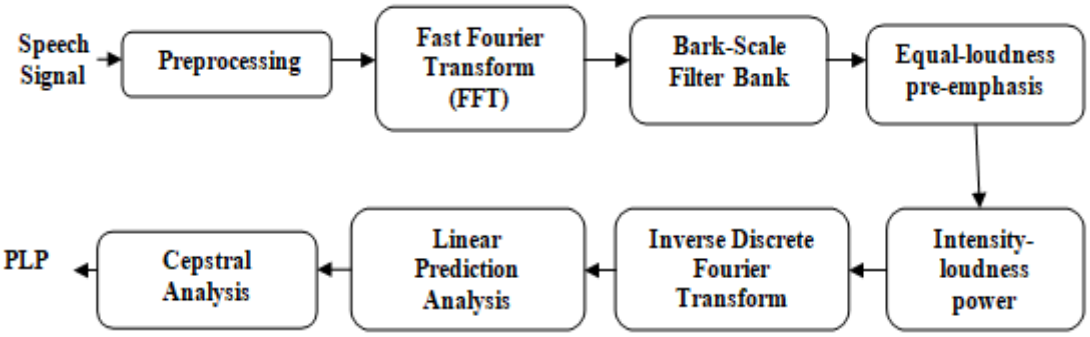

Figure 11: PLP Processor

These feature extraction approaches enable speech recognition by extracting useful information from voice signals. The speech signal of short period of time ranging 5 to 100msec is adequately short. For short time spectral analysis the feature extraction technique like MFCC, LPCC and PLP are used normally. The serious challenge faced during speech recognition is the noise. In recent years many robust speech recognition technique has been studied.

2.2 Visual Feature Extraction

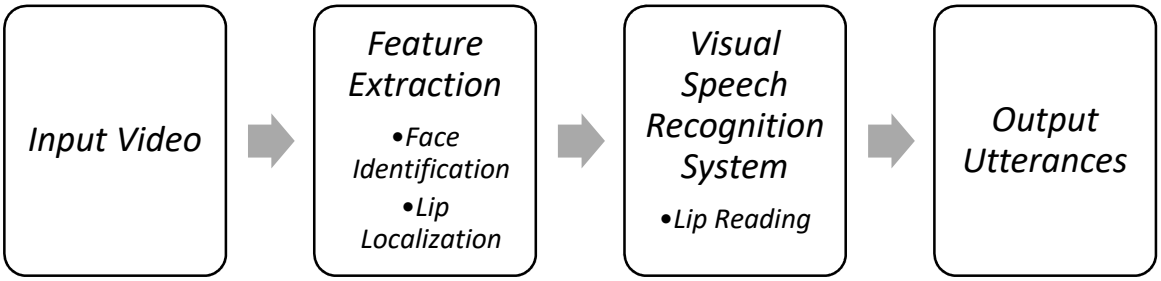

Figure 12: Workflow of VSR System

Figure 12, shows the various parts of 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 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 colour in a simpler way. Image based approach includes RGBModel, Hue Saturation Value Model and YCbcrcolor model.
- Model based approach: Active Shape Modeling (ASM) and Active Appearance Model (AAM) are the extensively used model based approaches. In ASM model each object is represented by a landmark which is a chain of trait points as shown in figure 13. Figure 14 shows the lip localization using height width ratio method.

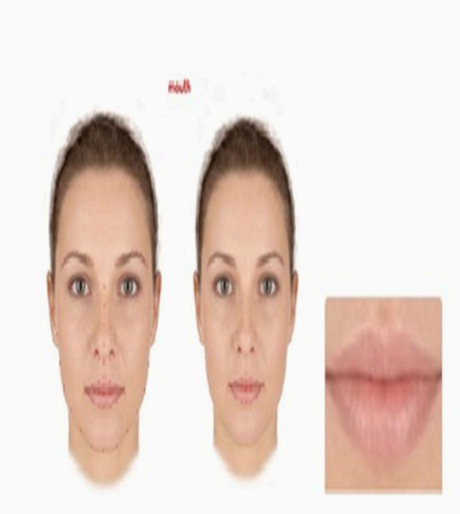

Figure 13: Active Shape Modeling

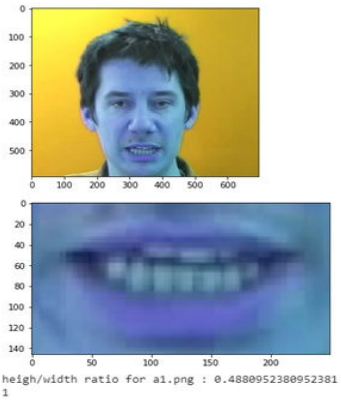

Figure 14: Lip Localization using height Width Ratio

Lip localization can also be done using 68 shape predictor method, it is publicly available through dlib package. It utilizes 68 landmark points to identify the lip region as given in figure 15.

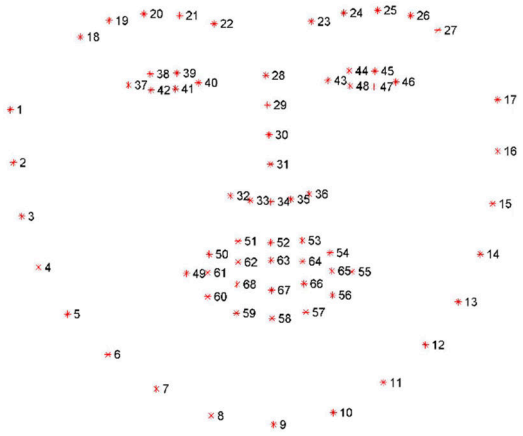

Figure 15: 68 Shape Predictor Method

2.3 Benchmark Dataset Statistic on Audio Visual Speech Recognition

Table 1 shows the benchmark Audio Visual Speech dataset used in recent studies. Table 2 shows the different existing models developed on AVSR.

Table 1: Dataset Statistic on AVSR

	Recording Nature	**No of Speakers**	**Language**	**Isolated /continuous Speech**	**Words**	**Sentences**
GRID	Recorded in studio environment	34 (18 male, 16 female)	English	Isloated	165,000	
TCD Timit	Recorded in studio environment	62	English	Continuous		6913 Sentences
AVletter	Recorded in studio environment	10	English	Isloated	780 Words	
AVletter2	Recorded in studio environment	5	English	Isloated		
LRS-TED	TED and TEDx videos	15088	English	Continuous	807,375	

Table 2: AVSR existing Models

Model	**Dataset used**	**Techniques used**	**WER obtained**
Lip Net[8]	GRID	RNN-CTC	4.9
Lip Type[9]	GRID	3D CNN + Bi-GRU +5-gram Language model	2.6
Deep Lip Reading[10]	LRS	Bi LSTM, Transformer based model	-

3 Conclusion

Speech recognition system requires high accurate feature extraction techniques because there is only minute difference in the lip movement for each phoneme. Combining audio and visual features will give additional information for the speech recognition system to identify the spoken utterances under various circumstances. The various feature extraction techniques are detailed in this paper which would be useful for the future research work carried out in Audio Visual Speech Recognition systems.

References

[1] Singh, A., Kadyan, V., Kumar, M. et al. (2020). ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. ArtifIntell Rev 53, 3673–704.https://doi.org/10.1007/s10462-019-09775-8

[2] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4960-4964, doi: 10.1109/ICASSP.2016.7472621.

[3] D. S. Vijayan, A. Mohan, J. J. Daniel, V. Gokulnath, B. Saravanan, and P. D. Kumar, "Experimental Investigation on the Ecofriendly External Wrapping of Glass Fiber Reinforced Polymer in Concrete Columns," vol. 2021, 2021.

[4] AmirsinaTorfiet. al,"3D Convolutional Neural Networks for Cross Audio Matching Recognition", IEEE, October 2017[DOI0.1109/ACCESS.2017.2761539]

[5] FatemehVakhshitehet.al.,"Lip reading via deep neural networks using hybrid visual features ",Image Anal stereol,pages:159-171,Doi:10.556/ias.1859,2018.

[6] Guoyingzhao et .al.,"Lip reading with local spatiotemporal Descriptor", IEEE transaction on Multimedia ,11(7):1254- 1265,2009.

[7] TriantafyllosAfouraset. al. , "Deep audio-Visual speech recognition", Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi/org/10.1109 /TPAMI.2018.2889502.

[8] YannisM.Assael, "Lipnet END-TO-END Sentence- level LIP reading" 16 Dec2016,arXiv:1611.0159.

[9] V. L. Shruthi, M. Kalpana and D. S. Vijayan, An experimental study on mechanical properties of bitumen added with industrial waste steel stag, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2020.03.032

[10] Afouras, T. and Chung, J.~S. and Zisserman, A. "Deep Lip Reading: a comparison of models and an online application" InterSpeech, 2018.