IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AUDIO VISUAL SPEECH RECOGNITION FOR HEARING IMPAIRED CHILDREN

Srikanth S, Asst Prof. SVCE, Bengaluru India, Namitha C, SVCE Bengaluru India, Nanditha N E, SVCE Bengaluru India, Nasim Akhtar, SVCE Bengaluru India

Abstract: Assistive technology holds tremendous promise in aiding individuals with hearing impairments through the utilization of Audio-Visual Speech Recognition (AVSR). Approximately 466 million people globally contend with hearing loss, with hearing-impaired individuals often relying on lip reading to comprehend speech. Challenges such as a shortage of trained sign language interpreters and the prohibitive costs of assistive devices further compound the difficulties faced by these individuals. To bridge these gaps, our research endeavours to leverage advanced deep learning models to develop a novel visual speech recognition technique. Unlike existing VSR methods, which suffer from inaccuracies, our approach involves integrating outcomes from both audio and visual speech modalities. Our study introduces a cutting-edge deep learning-based audio-visual speech recognition model tailored for efficient lip reading. Through our efforts, we aim to enhance system performance, achieving a notable reduction in word error rate to approximately 6.59% for the ASR system and a lip-reading model accuracy of around 95%.

I. INTRODUCTION

The realm of Speech recognition spans various disciplines, including natural language processing, signal processing, and artificial intelligence. Speech, characterized by a continuous stream of sound comprising phonemes, serves as the primary mode of human interaction, with hearing-impaired individuals relying on lip reading to understand spoken words. Fig. 1 depicts the current landscape of assistive technology for hearing-impaired students. Sign language remains a widely adopted practice among these students for learning, yet challenges persist from both the perspectives of facilitators and students.

From the teacher's viewpoint:

- Shortage of adequately trained sign language instructors.
- Limited awareness regarding emerging technologies.

From the student's viewpoint:

- Limited availability of user-friendly assistive devices.
- Challenges in interpreting information.
- Excessive costs associated with acquiring assistive devices.
- Insufficient e-learning materials with subtitles.

To address the hurdles prevalent in the current assistive technology landscape, this paper proposes the use of audio-visual speech recognition to enhance lip reading accuracy. Audio speech recognition involves the automated conversion of audio features into text.

Datasets such as Liberises and Timit are commonly utilized for automatic speech recognition tasks. Deep learning architectures like Deep Speech, LAS, and Wav2Letter have demonstrated superior recognition performance in this domain. Visual Speech Recognition (VSR) systems have gained prominence in recent years due to their independence from acoustic environments. These systems automatically detect spoken words by tracking the movements of the speaker's lips, offering an alternative mode of communication for individuals with hearing impairments. VSR plays a crucial role in Audio Visual Speech Recognition Systems (AVSR), enhancing the performance of audio-based recognition systems. VSR technologies are now being applied in noisy outdoor settings such as driving or mobile phone conversations. Various approaches are being explored to improve VSR system performance, catering to diverse real-world applications.

II. RELATED WORK

Lip reading involves discerning spoken words through the observation of lip movements, while Visual Speech Recognition (VSR) automates this process. Phonemes serve as the foundational linguistic units, whereas Visemes are the basic visual units utilized in lip reading systems. Despite advancements, hearing-impaired individuals have historically struggled with word identification (Easton & Basala, 1982; Fisher, 1968). The pioneering work of Petagan et al. (Petajan, Bischoff, Bodoff & Brooke, 1988) introduced the first VSR method based on the height-weight ratio.

Subsequent research has concentrated on enhancing automatic lip reading through visual speech recognition techniques (Torfi, Iranmanesh, Nasrabadi & Dawson, 2017; Vakhshiteh, Almasganj & Nickabadi, 2018).

Previous VSR methodologies encompassed features such as mutual knowledge, quality, tongue appearance, and teeth appearance (Heckmann, Savariaux, Berthommier & Frédéric, 2002). Machine learning classifiers like Support Vector Machines (Gordan, Kotropoulos & Pitas, 2002; Frolov & Sadykhov, 2009) and hidden Markov models (Puviarasan & Palanivel, 2011) have been employed for lip image classification. Advanced deep learning models like Convolutional Neural Networks (CNNs) have found widespread application in fields like facial recognition and medical diagnostics (Wang et al., 2021; Zhang, Satapathy, Guttery, Gorriz & Wang, 2021).

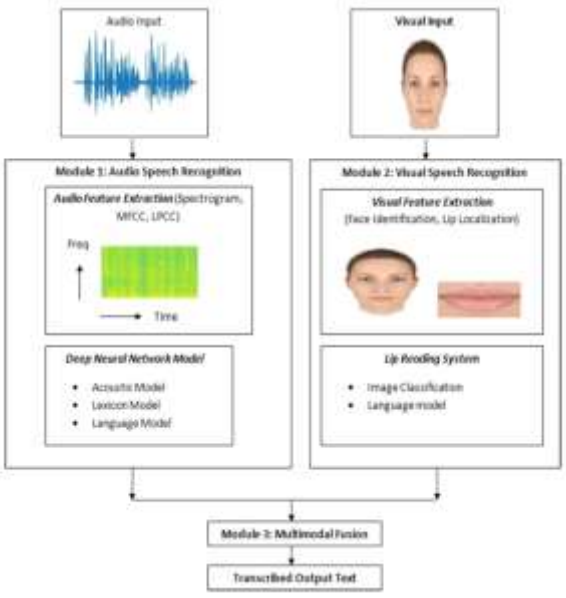
Recent research has emphasized the superiority of deep learning algorithms (Alothmany, Boston, Li, Shaiman & Durrant, 2010; Feng, Guan, Li, Zhang & Luo, 2017) over traditional machine learning approaches in speech analytics. Oxford University's development of lipnet (Assael, Shillingford, Whiteson & Freitas, 2016) and its enhanced version, liptype (Pandey & Arif, 2021), represents significant progress in this domain.

Automatic Speech Recognition (ASR) systems map variable-length audio signals to

sequences of words. While traditional methods such as Hidden Markov Models (HMMs) and Gaussian Markov Models (GMMs) have been employed, they often exhibit higher word error rates.

Zeghidour et al. (2018) demonstrated that noise conditions adversely affect recognition models. Hybrid models combining Long Short-Term Memory (LSTM)-based language models with HMM-based speech recognition systems have achieved low word error rates (Zhou, Schlu¨ter & Ney, 2020). Listen Attend and Spell (LAS), a combination of recurrent neural network encoders and attention-based decoders, has achieved promising results (Rose, Kumar & Renuka, 2019).

Numerous studies (G, A, K, D, & Karthika, 2020; MC, Renuka & Kumar, 2021) have explored speech analytics using various deep learning techniques across diverse research applications.



III. METHODOLOGY

In the domain of Audio-Visual Speech Recognition (AVSR), there exist three distinct modules: audio speech recognition, visual speech recognition, and multimodal fusion. The fundamental architecture of audio-visual speech recognition comprises these components.

Audio speech recognition refers to the process of transcribing spoken utterances into text. To train the neural network model, the Librispeech dataset is commonly utilized. Feature extraction is then applied to capture pertinent information from the audio input. Various representation methods are employed to convert the one-dimensional input speech signal into two-dimensional representations. Among these, Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms are prevalent. MFCC features are computed by first determining the logarithm of the Mel frequency using Equation (1), where the frequency (mel) is calculated as a function of the frequency in Hertz (Hz). Subsequently, MFCC values are derived using Equation (2), involving the application of discrete cosine transform (DCT) to the computed Mel frequency values.

$$\text{mel}(f) = 2595 * \log(1 + \frac{f}{100}) \quad (1)$$

$$cn = \sum_{n=1}^{k} \log S k \cos[\frac{n(k-\frac{1}{2})\pi}{k}] \quad (2)$$

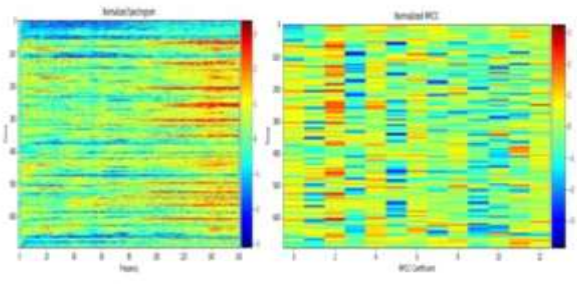


Fig. 3. Spectrogram and MFCC

IV. 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 Modelled?	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

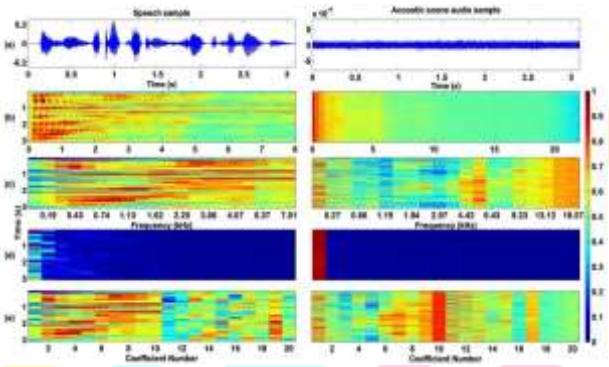


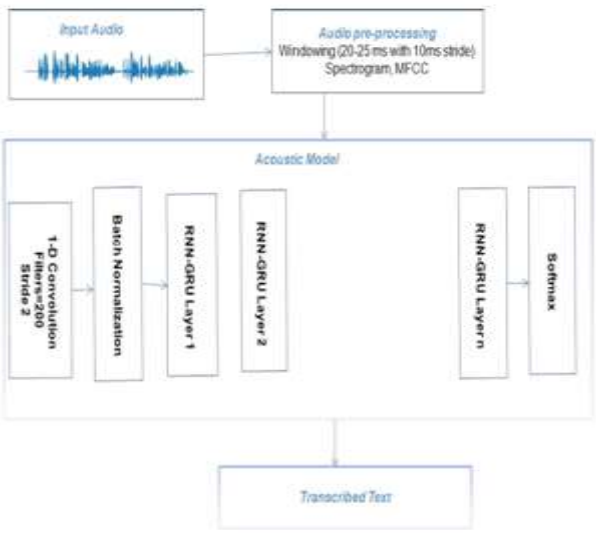
Table 2
Pseudo code for AVSR System.

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, …x n} is fed into stack of GRU layer with softmax activation at end. Step 3: The output characters {y 1, y 2, 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.

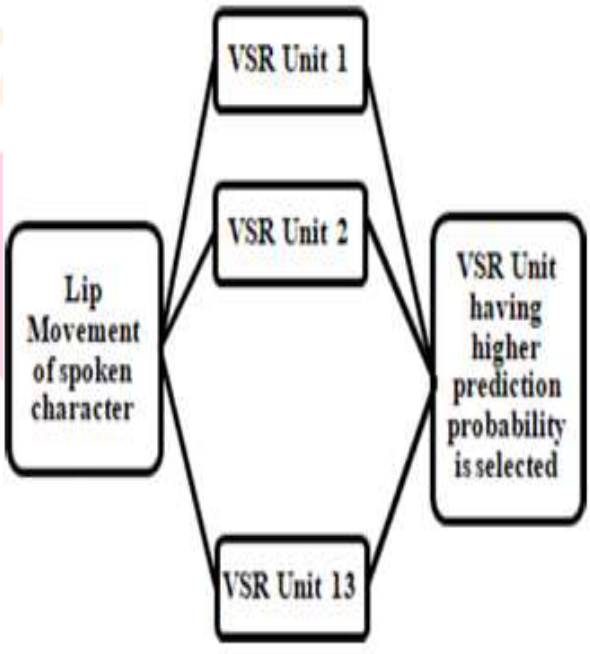
Table 3
Prediction Probability Algorithm.

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.



V. Visual speech recognition (visemes to text)

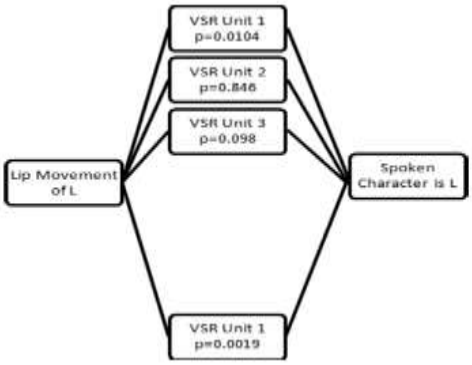
Visual speech recognition (VSR) is an automated process that deciphers spoken words through analysis of lip movements, offering a valuable communication avenue for individuals with hearing impairments. The dataset used for training can be generated at either the character or word level. Character-level datasets are composed of individual letters or phonemes, while word-level datasets focus on specific words. Given the complexity of training on entire words, the character model dataset from the GRID corpus is preferred for training the proposed neural network model.



The GRID dataset, curated by Cooke in a controlled recording environment, serves as a robust foundation for this technology. The VSR system comprises several components,

including Face Identification, Lip Localization, and Lip-Reading System.

Face identification involves recognizing faces within images, achievable through traditional methods or deep learning approaches. Lip localization entails pinpointing the lip region or the Region of Interest (ROI) within identified faces, employing either image-based or model-based approaches.



In the proposed architecture, videos are segmented into frames, with unique frames fed into the VSR unit. For instance, a video uttering "add" would yield distinct frames for each phoneme ("a," "d," "d"). These frames are inputted into the VSR system to predict the spoken characters.

The prediction probability algorithm for sampling the dataset is outlined, ensuring efficient character recognition. Each VSR unit, comprising two character models, employs Active Shape Modeling (ASM) to extract lip movements.

ASM detects facial landmarks and adjusts them to the correct positions, representing the mouth region with 48–68 trait points. The extracted lip movements are then fed into each VSR unit, with prediction probabilities calculated independently. The output of the VSR unit with the highest prediction probability determines the model's output.

The VSR unit is trained with two characters.

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

The spoken character is given to each VSR unit and the probability prediction is calculated as shown below, Probability prediction = $P\ (\ X,\ C\ )$ 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 :$Y\ (\ X\ ) = max\ [\ P\ (\ V\ SR\ 1\ )\ ,\ P\ (\ V\ SR\ 2\ )\ ,\ P\ (\ V\ SR\ 13\ )\ ]$ where, k is the number of melcepstrum coefficients, Sk is the output of filter bank and Cn is the final MFCC coefficients.

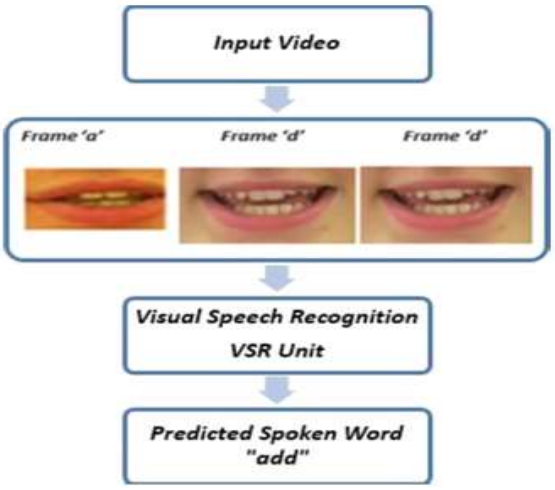


Table 4
Dataset Description.

Dataset	Description
Libri Speech	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.

The experimental analysis utilizes two distinct datasets: LibriSpeech and Grid, as detailed in Table 4 and Table 5.

Model training is conducted on a high-performance workstation equipped with a GeForce RTX card, leveraging the TensorFlow framework.

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

Alphabets are extracted from the GRID dataset, and the ASM model is employed to isolate the lip region. Lip movement data (L) is fed into all 13 VSR units, comprising stacked CNNs and max-pooling layers. Prediction probabilities are computed at each VSR unit, with VSR unit 6 demonstrating the highest prediction probability. The Word Error Rate (WER), a crucial evaluation metric for speech recognition, quantifies the ratio of errors to the total number of words.

WER is calculated using the Levenshtein distance formula, depicted. In this study, the RNN-GRU speech-to-text model is juxtaposed against architectures like wav2letter+ and DeepSpeech2, as outlined in Table 6.

Table 6

Comparison of Speech to Text Models.

Models	Layers Used	WER%
Wav2 Letter [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	6.59

CONCLUSION

The advancement of an Enhanced Audio-Visual Speech Recognition (AVSR) system holds immense potential as an assistive technology for individuals with hearing impairments and for detecting speech in noisy environments.

Our proposed framework leverages deep learning algorithms to construct an efficient model. Lip localization is achieved using ASM, while a CNN-based VSR unit is incorporated to enhance overall performance.

The system achieves a remarkable accuracy of approximately 95%, coupled with a reduced word error rate of 6.59%. Integrating visual information with speech recognition through deep learning algorithms presents a highly effective visual speech recognition system.

VSR serves as a pivotal tool for speech recognition in the absence of audio, and it complements Audio-Visual Speech Recognition (AVSR) systems.

Looking ahead, our future endeavors include proposing a BERT-based language model to further enhance the Word Error Rate (WER) of our audio model.

Additionally, we aim to develop an audio-visual multimodal fusion framework to enhance the performance of automatic speech recognition.

REFERENCES

1. Suhm,B.,Neuschaefer-Rube,C., Nuernberger, B., & Schmidt, R. (2012). Audio-visual speech recognition for hearing-impaired listeners in reverberation. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2012, 739-742.
2. Ding, J., Huang, W., & Hao, J. (2019). Deep Audio-Visual Speech Recognition for Hearing Impaired Users. Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2019, 2808-2812.
3. Zhang, S., Zhang, Y., Li, H., & Dai, L.R. (2020). A Deep Learning Approach to Audio-Visual Speech Recognition for Hearing Impaired People. IEEE Access, 8, 188127-188136.
4. Wang, Y., Liu, H., & Zhang, J. (2017). Audio-visual speech recognition for hearing-impaired listeners based on

auditory saliency and lip movements. Applied Acoustics, 116, 133-139.

5. Wu, Z., Ruan, J., Lin, X., Jiang, Y., & Dai, L.R. (2017). Enhancing Audio-Visual Speech Recognition by Cross-Modal Distillation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, 5600-5604.
6. Deng, J., Zhang, X., Liu, D., & Wu, Q. (2020). A Deep Learning Framework for Audio-Visual Speech Recognition Considering Multiple Contexts. IEEE Access, 8, 190787-190799.