Visual Speech Recognition Using Lips Movement

Dr. A. B. Gavali[1], Pooja Shankar Ghugarkar[2], Supriya Raghunath Khatake[3], Rajanandini Anil Kothawale[4]

1,2,3,4 S. B. Patil College of Engineering, Indapur, Computer Engineering

---

Abstract - The audio-visual speech recognition technology seeks to increase noise-robustness in mobile contexts by extracting lip movement from side-face pictures. Prior bimodal speech recognition algorithms depended on frontal face (lip) photos, but because users must be able to speak while holding a device in front of their face, these techniques are difficult for users to utilise. Our suggested method records lip movement using a tiny camera built into a smartphone, making it more practical, simple, and natural. Additionally, this technique effectively prevents a decline in the input voice's signal-to-noise ratio. (SNR). For CNN-based recognition, visual properties are extracted via optical-flow analysis and merged with audio data.

***Key Words*:** Convolutional Neural Network, Deep Learning, Image.

1.INTRODUCTION

Since speech is fundamental, simple, and everyone can use it without a device or any technological knowledge, it is an essential part of communication. The problem with crude connecting devices is that they require some fundamental technical know-how to operate. As a result, non-technical users may have trouble using these devices to communicate. Users will benefit from speaking to computers in their native tongues rather than giving input from external devices because speech recognition is the main focus of this work and no technical knowledge is required. Computer usage, including issues with usability and effectiveness of computer interface, is the focus of today's typical technical concerns. English literature knowledge is now practically required in order to utilise computers and other modern technology. This makes it difficult for most people to use computers and other technology. Regular people must stay up with these advancements since information technology is advancing swiftly.

A more user-friendly method would need to be devised in addition to this restriction. The most user-friendly system, for instance, would have components that can read input as speech in regional languages, accept it, and respond to those regional cues. This enables common people to benefit from technology advancements. The complimentary qualities offered by the visual information cannot be reduced by background noise. Because they are well understood, acoustic features are used for speech recognition. The decision of visual characteristics, the fusion model for visual and aural input, and the recognizer represent the main hurdles. The visual characteristics are the most crucial idea in VSR. (Visual speech recognition). Any auditory disturbance or interruptions in a noisy environment have no effect on this. Intriguing research on visual speech has helped develop digital entertainment, security, and human-computer interaction, among other things. As a result, the suggested method for speech recognition only makes use of visual traits. Because of the facts, academics have carried out specific VSR (visual speech recognition) and AVSR (audiovisual speech recognition) investigations. (audio-visual speech recognition). In visual speech recognition, this method is known as the automated lip-reading strategy. Numerous automatic speech recognition systems that use both audio and visual data have been launched in recent years. In all of these kinds of systems, one of the key goals of visual speech recognizers is to improve identification accuracy, particularly in noisy environments. This study's main objective is to use lip features to create VSR (visual speech recognition) for Indian languages. The idea behind this study is to select an input video that has all pertinent background information and lighting conditions, then extract the output text from it. GLCM (Gray Level Cooccurrence Matrix), Gabor convolve, and canny edge detection which is particularly effective at locating the lip edge are just a few of the methods used to extract the shape and texture of lips. The output may then be classified using a CNN classifier based on the feature vector.

2. PROPOSED SYSTEM

2.1 PROBLEM STATEMENT:

To Predict the text based on the lips movement of individual person also gives the output in audio format.

2.2 EXISTING SYSYEM:

1. Low accuracy detected.
2. Audio Output is not predicted.

2.3 SYSTEM ARCHITECTURE:

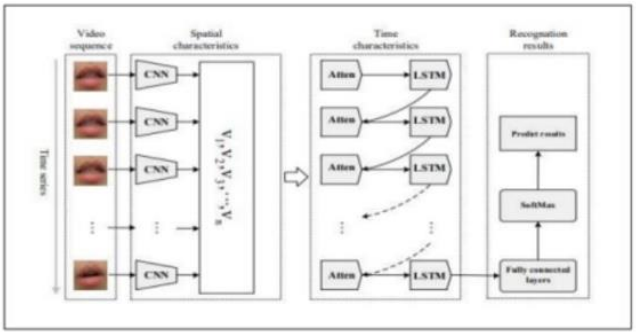

Fig. 1 Architecture

2.4 MATHEMATICAL MODULE:

Let S be as system which allow users to recognize the Visual speech recognition using lip movement for deaf people using Deep Learning. S = In, P, Op,

Identify Input In as

In =Q

Where,

Q = Input Video Identify Process P as

P = CB, C, PR

Where,

CB = Perform preprocessing on the input

C = Extractions of features and storing extracted features for further comparison

PR = Classification using Convolutional Neural Network Identify Output Op as

Op = UB

Where,

UB = Updated Result

Failures: Huge database can lead to more time consumption to get the information. Hardware failure Software failure.

Success: Search the required information from available in datasets. User gets result very fast according to their needs.

Time Complexity: Check number of patterns available in the dataset=n If(n,1) then retrieving of information can be time consuming. So, the time complexity of this algorithm is O(n^n).

3. SYSTEM REQUIREMENTS

3.1 SOFTWARE REQUIREMENTS:

- Operating system: windows
- Coding Language: Html, Python 2.7 onwards
- Database: MYSQL
- IDE: Python
- CNN deep learning package

3.2 HARDWARE REQUIREMENTS:

- Processor: - Intel Pentium 4 or above
- Memory: - 2 GB or above
- Other peripheral: - Printer
- Hard Disk: - 500gb

3.3 ADVANTAGES:

- Can be used for bio-metric identification in banking system.
- Can be a useful system for mute people.
- Can be useful in spying activities.

3.4 APPLICATION:

- Bio-metric identification.
- Spying activities.
- Useful for mute people.

4. CONCLUSIONS

According to a recent study, recurrent neural networks are being utilized to try to find the optimal temporal sequence. Because CNN may store both short- and long-term context information in their cell, they are widely used to simulate sequences, however it is unclear how to

best make use of this ability. Some authors, for example, have used numerous CNN layers to simulate various scales of context in order to include restrictions relevant to larger speech structures, such as connected phonemes, syllables, phrases, or sentences.

5. REFERENCES

1. Zhang, Xingxuan, Feng Cheng, and Shilin Wang." Spatio-temporal fusion based convolutional sequence learning for lip reading." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
2. Kurniawan, Adriana, and Suyanto Suyanto." Syllable-Based Indonesian Lip-Reading Model." 2020 8th International Conference on Information and Communication Technology (ICoICT). IEEE, 2020.
3. Michelsanti, Daniel, et al." An overview of deep-learning-based audio-visual speech enhancement and separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).
4. Desai, Dhairya, et al." Visual Speech Recognition." International Journal of Engineering Research Technology (IJERT) 9.04 (2020).
5. Fenghour, Souheil, et al." Deep Learning-based Automated Lip-Reading: A Survey." IEEE Access (2021).
6. Afouras, Triantafyllos, et al." Deep audio-visual speech recognition." IEEE transactions on pattern analysis and machine intelligence (2018).
7. Zhang, Yuanhang, et al." Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.
8. Petridis, Stavros, et al." End-to-end visual speech recognition for small-scale datasets." Pattern Recognition Letters 131 (2020): 421-427.
9. Lee, Wookey, et al." Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review." Sensors 21.4 (2021): 1399.
10. Lee, Yong-Hyeok, et al." Audio–visual speech recognition based on dual crossmodality attentions with the transformer model." Applied Sciences 10.20 (2020): 7263.