IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

Visual Speech Recognition

Supriya Patil[1], Vaibhav Dhoble[2], Saatvik Gawade[3], Pratiksha Jagdale[4], Rohan Jinde[5]
Assistant Professor, Department of Information Technology[1]
Student, Department of Information Technology[2,3,4,5]
Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract: *The audio-visual speech recognition approach attempts to boost noise-robustness in mobile situations by extracting lip movement from side-face images. Although earlier bimodal speech recognition algorithms used frontal face (lip) images, these approaches are difficult for consumers to utilize because they need them to talk while holding a device with a camera in front of their face. Our proposed solution, which uses a small camera put in a handset to capture lip movement, is more natural, simple, and convenient. This approach also effectively avoids a reduction in the input speech's signal-to-noise ratio (SNR). Optical-flow analysis extracts visual features, which are then coupled with audio features in the context of CNN-based recognition.*

Keywords: Convolutional Neural Network, Deep Learning, Image

I. INTRODUCTION

Speech is a key criterion for communication since it is easy, simple, and everyone may speak without the use of any technology, and it does not require a high level of technical knowledge. The problem with primitive interfacing devices is that they require a certain amount of fundamental skill set to operate. As a result, interacting with such devices will be challenging for those who lack technological knowledge. Because the major focus of this work is on speech recognition and no technical skills are necessary, it will be beneficial for people to speak to computers in familiar language rather than giving inputs from other systems' devices.

Nowadays, typical technological challenges revolve around computer usage, such as how effective computer interaction is and how user-friendly less conventional ways are. Knowing English literature has practically become a need for interacting with computers and using information technologies. This makes it difficult for ordinary people to use computers and other electronic gadgets. Because information technology is rapidly improving, it is critical for ordinary people to stay on track with technical advancements.

Aside from this restriction, a more approachable system will need to be built, such as devices that can read and receive input as regional language speech and respond to those regional things for the most user-friendly system. This enables ordinary people to benefit from technical advancements. The complementing qualities offered by the visual information cannot be corrupted by audio noise in the environment. Because acoustic properties are well recognized d, they are used for speech recognition. The selection of visual features, the fusion model for visual and audio data, and the recognizer are the main difficulties. The visual parameters are the most significant idea in VSR (visual speech recognition). Any acoustic noise or disturbances in a noisy environment will have no effect on this. Visual speech is a fascinating study issue that has been applied to a variety of domains, including improving human-computer interaction, security, and digital entertainment. As a result, the suggested methodology focuses solely on visual features to recognize speech. The facts have prompted academics to conduct specific VSR (visual speech recognition) and AVSR (audiovisual speech recognition) studies (audio-visual speech recognition). For visual speech recognition, this is known as the automatic lip-reading approach. Several automatic speech recognition algorithms that integrate both audio and visual data have been presented in recent years. An key goal of visual speech recognizers in all of these types of systems is to enhance recognition accuracy, especially in noisy environments. The main focus of this work is on VSR (visual speech recognition) for Indian languages using lip characteristics, and the entire concept will be based on the selection of input video with all light and ambient conditions, followed by the extraction of text output. Many techniques, such as canny edge detection, which is particularly useful for recognizing the lips edge, GLCM (Gray Level Cooccurrence Matrix), and Gabor convolve, are used to extract the form and texture aspects of lips. Finally, the output can be categorized using a CNN classifier based on the feature vector.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

II. LITERATURE SURVEY

Convolutional sequence learning based on spatio-temporal fusion for lip reading [1]. A Temporal Focal block to accurately depict short-range relationships, as well as a Spatio-Temporal Fusion Module (STFM) to maintain local spatial information while lowering feature dimensions. As indicated by the experiment results, our solution delivers equal performance to the state-of-the-art techniques while using substantially less training data and a much lighter Convolutional Feature Extractor.

Indonesian Lip Reading Model Based on Syllables [2]. You can construct a new phrase that isn't in the dictionary using the syllable-based paradigm. By mixing the syllables that already exist, a new word is generated. Because the data acquired is too small for deep learning, the augmentation step is repeated 40 times.

An overview of audio-visual speech augmentation and separation based on deep learning [3]. A comprehensive examination of this field of study, focusing on the major elements that separate systems in the literature: audio features, visual features, deep learning methods, fusion approaches, training objectives, and objective functions. Since they may employ these techniques to better and separate audio-visual speech, deep-learning-based approaches for voice reconstruction from silent movies and audio-visual sound source separation for non-speech data are being investigated.

Visual Speech Recognition (VSR) is a technique for recognizing speech using images. The following is an overview of numerous Machine Learning algorithms and image processing processes for efficiently extracting and tracking lip movements. Image processing is now a common method for extracting important traits and using numerous environmental aspects to improve the end product. The paper's main focus is on a comparison of many VSR algorithms. Categorization methods include LSTMs, CNNs, Decision Trees, and Neural Networks, to name a few.

Based on deep learning A Survey on Automated Lip-Reading [5]. Audio-visual databases, feature extraction, classification networks, and classification schemas are all components of automated lip-reading systems that are compared. The field of automated lip-reading research is vast. Because of advances in deep neural networks and the introduction of large-scale databases containing vocabularies with thousands of distinct words, lip-reading algorithms have gone from recognizing solitary speech units in the form of numbers and letters to decoding complete sentences.

Deep Audio-Visual Speech Recognition [6] is a technique for recognizing speech in both audio and visual formats. LRS2-BBC is a large-scale, unconstrained audiovisual dataset made up of thousands of movies collected and preprocessed from British television. On the LRS2-BBC lip reading dataset, the top visual-only model outperforms the prior state-of-the-art by a considerable margin and establishes a strong baseline for the recently released LRS3-TED. Finally, we show that even when a clear audio stream is available, visual information can help boost speech recognition accuracy. Combining the two modalities improves performance significantly, especially when there is noise in the audio.

Is it possible to read speech without looking at the lips? Rethinking RoI Selection for Deep Visual Speech Recognition [7] a comprehensive study using state-of-the-art VSR models to assess the effects of several facial regions, including the lips, the entire face, the upper face, and even the cheeks. Experiments are carried out on benchmarks with diverse properties at the word and sentence levels. Incorporating information from extraoral facial regions, including the upper face, reliably improves VSR performance, despite the data's complicated fluctuations. In addition, we present a simple but successful strategy based on Cutout for learning new discriminative features for face based VSR, with the goal of maximizing the utility of information stored in various facial areas.

Visual Speech Recognition for Small-Scale Datasets (End-to-End) [8]. An end-to-end visual speech recognition system based on fully connected layers and Long-Short Memory (LSTM) networks for small-scale datasets is described. The model is divided into two streams: one that extracts features directly from mouth photos, and the other that derives features from difference images. The temporal dynamics in each stream are modelled using a Bidirectional LSTM (BLSTM), which is then combined using another BLSTM. The proposed model achieves state-of-the-art performance on all four datasets, OuluVS2, CUAVE, AVLetters, and AVLetters2, greatly exceeding all previous approaches published in the literature, including CNNs pre-trained on external databases.

A Review of Biosignal Sensors and Deep Learning-Based Speech Recognition The interface technologies, which are mouth-mounted devices for speech recognition, production, and volitional control, and the corresponding research to develop artificial mouth technologies based on various sensors, such as electromyography (EMG), electroencephalography (EEG), electropalatography (EPG), electromagnetic articulography (EMA), permanent magnet articulography (PMA), gyros, images, and three-dimensional magnetic sensors, especially with deep learning techniques. We investigate a variety of deep learning technologies linked to voice recognition, such as visual speech recognition and silent speech interface, as well as

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

their flow and classification into a taxonomy. Finally, we explore approaches for resolving communication challenges in people with impairments who have difficulty communicating, as well as future research on deep learning components.

With the Transformer Model, audio–visual speech recognition is based on dual cross-modality attentions [10]. an AVSR model with DCM attention and a hybrid CTC/attention architecture based on the transformer We used a hybrid CTC/attention structure to improve monotonic alignments and built the DCM attention for correct alignment information between audio and visual modality even with noisy reverberant audio data. In general, our model outperformed the transformer-based models in terms of recognition, even for out-of-sync input, and the hybrid CTC/attention loss further improved the performance.

III. Objectives of System

- To build and implement a deep learning-based solution for text audio detection utilising lips movements.
- Convolutions Neural Network was used to extract several lips movement futures and identify the voice.
- Develop a method for improving the accuracy of voice recognition at the word and sentence levels.
- To compare and contrast the suggested system's results with those of other systems.

IV. Implementation Details Of Module

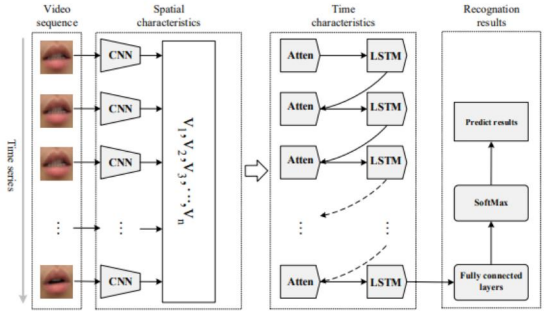

Figure a: Block Diagram

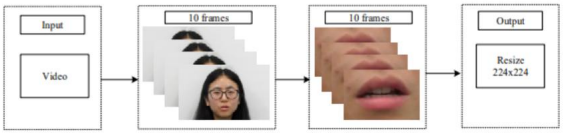

Figure b: System Steps

The proposed structure and essential steps are addressed in depth in the four parts that follow. The dynamic lip videos must first be preprocessed, which includes separating audio and video signals, extracting keyframes, and positioning the mouth. Second, CNN is used to extract features from the preprocessed picture dataset.

Then, to learn sequence information and attention weights, we combine LSTM with an attention mechanism. Finally, the ten-dimensional characteristics are mapped using two completely linked layers, with the SoftMax layer predicting the result of automatic lip-reading recognition. SoftMax normalizes and categorizes the output of fully linked layers based on likelihood.

V. Conclusion

Recent research suggests that the best modelling of temporal sequences is still an unresolved subject that is being addressed with recurrent neural networks.

Because of their ability to preserve both short- and long-term context information in their cell architectures, CNN have been widely utilized for modelling sequences, albeit it is unclear how to fully use this feature. Several authors, for example, have used numerous CNN layers to mimic different scales of context, with the goal of introducing constraints relating to larger speech structures such as connected phonemes, syllables, phrases, or sentences.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, March 2022
Impact Factor: 6.252

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 Susanto. ”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.