Visual Speech Recognition Using Lip Movement for Deaf People Using Deep Learning

PROF . G. J. NAVALE[1], BHAVESH DALAL[2] NIKET PATIL[3], MAHIMA OSWAL[4], RIYA SATIJA[5]
[1, 2, 3, 4, 5] *Department of Computer Engineering*

Abstract— There has been a growing interest in creating automatic lip-reading systems (ALR). Methods based on Deep Learning (DL), like other computer vision applications, have grown in popularity and allowed for significant improvements in performance. The audio-visual speech recognition approach attempts to boost noise-robustness in mobile situations by extracting lip movement from side-face pictures. Although most earlier bimodal speech recognition algorithms used frontal face (lip) pictures, these approaches are difficult for consumers to utilise because they need them to speak while holding a device with a camera in front of their face. Our suggested solution, which uses a tiny camera put in a phone to capture lip movement, is more natural, simple, and convenient. This approach also successfully prevents a reduction in the input speech's signal-to-noise ratio (SNR). Optical-flow analysis extracts visual characteristics, which are then coupled with auditory data in the context of DCNN-based recognition. We employ DCNN for audio-visual speech recognition in this paper; specifically, we leverage deep learning from audio and visual characteristics for noise-resistant speech recognition. In the experimental analysis we achieved around 90% accuracy on real time test data that provides higher accuracy than traditional deep learning algorithm.

Indexed Terms-- Deep Learning, Image processing, clasisifcation, deep learning, feature extraction, feature selection

I. INTRODUCTION

Speech is a key criterion for communication since it is straightforward, simple, and everyone may talk without the use of any technology, and it does not need a high level of technical knowledge. The difficulty with primitive interfacing devices is that they need a certain amount of fundamental skill set to operate. As a result, interacting with such gadgets will be challenging for those who lack technological knowledge. Because the major focus of this work is on voice recognition and no technical skills are necessary, it will be beneficial for users to talk to computers in recognised languages rather than offering inputs from other systems' devices. Nowadays, typical technical challenges revolve around computer use, such as how effective computer interaction is and how user-friendly less traditional 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 technological gadgets. Because information technology is rapidly improving, it is critical for ordinary people to stay on track with technical advancements. Aside from this constraint, a more accessible system will need to be built, such as devices that can read and receive input as regional language speech and react to those regional things for the most user-friendly system. This enables ordinary people to benefit from technical advancements.

II. LITERATURE SURVEY

According to [1] a huge audio-visual (A/V) dataset of segmented utterances taken from public YouTube films, resulting in 31k hours of audio-visual training material On two large-vocabulary test sets, the performance of an audio-only, visual-only, and audio-visual system is compared: a collection of utterance segments from public YouTube videos called YTDEV18 and the publicly accessible LRS3-TED set. To emphasise the importance of the visual modality, we tested our system on the YTDEV18 set, which was purposely distorted with background noise and overlapping voice. While the performance of automated speech recognition (ASR) systems has

increased dramatically in recent years, there are still considerable difficulties for widespread ASR.

According to [2] Since the introduction of the attention mechanism in neural machine translation, attention has been integrated with or replaced the long short-term memory (LSTM) in a transformer model to address the LSTM's sequence-to-sequence (seq2seq) issues. Audio–visual speech recognition (AVSR), in contrast to neural machine translation, may increase performance by learning the association between audio and visual modalities. As a consequence, AVSR is difficult to train attentions with balanced modalities since the audio provides more information than the video connected to lips. We present a dual cross-modality (DCM) attention strategy that uses both an audio context vector using video query and a video context vector using audio query to raise the function of visual modality to that of audio modality by completely leveraging input information in learning attentions.

According to [3] A Parallel-WaveGAN-based scene classifier is used to create an efficient multiangle AVSR technique. The classifier determines whether voice data was captured in quiet or loud conditions. If our scene classification detects noisy settings, multi-angle AVSR is used to improve identification accuracy, however just ASR is used if the classifier predicts clear voice input to save processing time. We tested our approach with two multi-angle audio-visual databases: an English corpus with 5 views, OuluVS2, and a Japanese phrase corpus with 12 views, GAMVA.

According to [4] The approach entails recording a signal, identifying speech in it, recognising speech words in a simplified transcription, defining word boundaries, comparing a simplified transcription to a code book, and building a hypothesis regarding the degree of speech emotionality. When emotions are present, full identification of words and meanings of emotions happens in speech. The benefit of this approach is that it can be used by a large number of people since it does not need a lot of computer resources. When it comes to recognising good and negative emotions in a crowd, the mentioned approach may be used in public transportation, schools, and colleges, among other places.

According to [5] a deep stride convolutional neural network (DSCNN) architecture with artificial intelligence that uses the plain nets technique to learn prominent and discriminative features from spectrograms of voice signals that have been modified in previous phases to perform better. Local hidden patterns are learnt in convolutional layers rather than pooling layers, with particular steps to down-sample the feature maps, while global discriminative features are learned in fully connected layers. For the categorization of emotions in speech, a SoftMax classifier is utilised. On the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets, the suggested approach improves accuracy by 7.85 percent and 4.5 percent, respectively, while reducing model size by 34.5 MB.
According to [6] for practical applications in a variety of contexts, noise-resistant OCSR APIs based on an end-to-end lip-reading architecture are available. The Google Voice Command Dataset v2 was used to test the performance of many OCSR APIs, including Google, Microsoft, Amazon, and Naver. The Microsoft API was used with Google's trained word2vec model based on performance to provide more full semantic information for the keywords. For audio-visual speech recognition, the retrieved word vector was combined with the suggested lip-reading architecture. In the proposed lip-reading architecture, three types of convolutional neural networks (3D CNN, 3D dense connection CNN, and multilayer 3D CNN) were employed. After concatenation, the API and vision vectors were categorized.

According to [7] the purpose of this project is to recognise words and sentences said by a talking face, whether the audio is present or not. Unlike earlier research, which has focused on recognising a small number of words or phrases, we approach lip reading as an open-world issue with unconstrained natural language sentences and in-the-wild movies. LRS2-BBC and LRS3-TED are two large-scale, unconstrained audiovisual datasets created by gathering and preprocessing thousands of recordings from British television and YouTube. The first two models can transcribe audio and video speech sequences into characters, and they demonstrated that the same architectures can be employed when just one of the modalities is available.

According to [8] the goal of this research is to recognise words and phrases said by a talking face, regardless of whether or not audio is available. We approach lip reading as an open-world problem with unconstrained natural language sentences and in-the-wild videos, unlike previous research, which concentrated on identifying a limited number of words or phrases. The audiovisual datasets LRS2-BBC and LRS3-TED were developed by collecting and preprocessing hundreds of recordings from British television & YouTube. The first two models can transcribe audio or video speech sequences into characters, and they showed that when only one of the modalities is available, the same architectures may be used.

According to [9] train deep neural network-based A2V inversion models, a genuine 3D Audio-Visual Mandarin Continuous Speech (3DAV-MCS) corpus was employed. The inversion models' cross-domain adaptability enables adequate visual features to be created from audio input from mismatched domains. On two state-of-the-art Mandarin voice recognition tasks, DAPRA GALE broadcast transcription and BOLT conversational telephone speech recognition, the suggested crossdomain deep visual feature creation approaches were assessed. After both speaker adaptive training and sequence discriminative training, the AVSR systems built with cross-domain generated visual features consistently outperformed the baseline convolutional neural network (CNN) ASR systems by up to 3.3 % absolute (9.1% relative) character error rate (CER) reductions.

According to [10] the adversarial assaults are detected in a significant way by two audio-visual recognition models, LipReading in the Wild (LRW) and Geospatial Repository and Data (GRiD) Management, which are trained on Lip reading data sets. In comparison to Supervised Kernel Machines, Combined Neural Networks, and Band Feature Selection approaches, experimental findings show that the suggested strategy is a strong tool for identifying adversarial assaults. Over the past decade, deep learning algorithms have surmounted numerous multimedia analysis issues and performed better against classification and prediction tasks.

- Prposed system design

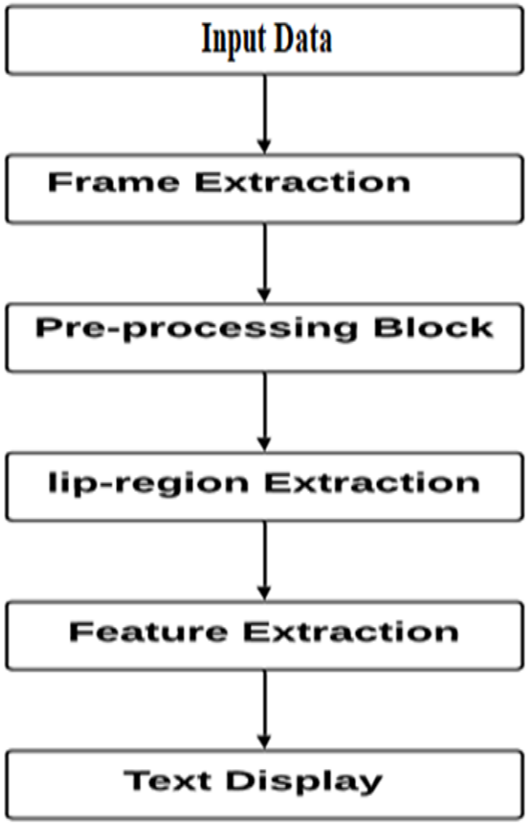

Figure 1 : Proposed system architecture

The below block diagram show the brief flow of our project where in it consist of seven blocks these are as follows: video input, frame extraction, pre-processing block, lip region extraction, feature extraction, discrimination analysis, text display.

Input:
This block is the input of our system, which contains the video of an individual's lip motion while speaking an alphabet. The video or image is captured so as to focus on the lip movement information. The video is taken under certain standardized conditions such as uniform illumination, steady face alignment. To build a model for implementing automated lip reading which involves Lip motion feature to text conversion?

Frame Extraction:
The video or image obtained is then converted into frames. The video can be converted into frames by method namely, Python.

Pre-processing:
Before the contour extraction algorithm is applied to the original image, it is required that it undergoes a sequence of transformations that eventually highlight the lip region. The preprocessing block performs such enhancements and color transforms on each frame image extracted.

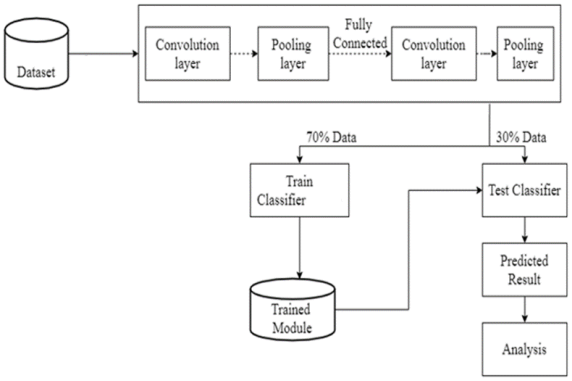

Figure 2: Logical workflow of proposed classifier (CNN)

Build a Classifier CNN
Deep Learning includes the Convolutional Neural Network (CNN). Too far, CNN has shown to be a highly efficient and successful method of achieving handwriting recognition. Convolutional Neural Networks are neural networks that employ multiple layers of filters to extract information from pictures.

1. Convolutional Layer: It's the foundation for constructing a CNN model. This layer conducted mathematical calculations on the picture that was used as input, as well as resizing the image into the M* M format. This layer's output describes the image's features, such as edge and corner mapping, also known as a feature map. The information was then added to the following layer.
2. Pooling Layer: This is the layer that connects the convolutional and fully connected layers. This layer is used to decrease the network's parameters and computation. The maxpooling and average pooling methods are provided by this layer. The most frequent method is max pooling. The output of the preceding layer, the pooling layer, is sent to the fully connected layer. This layer is where the categorization process takes place.
3. Classification and Recommendation: This layer test classifier has utilized for recommendation. The input features and trained modules has feed to test classifier and recommended the specific song for end user accordingly.

III. ALGORITHM DESIGN

1. Convolutional Neural Network (CNN)

Training

Input: Training dataset TrainData[], Various activation functions[], Threshold Th

Output: Extracted Features Feature_set[] for completed trained module.

Step 1: Set input block of data d[], activation function, epoch size,

Step 2 : Features.pkl ← ExtractFeatures(d[])

Step 3 : Feature_set[] ← optimized(Features.pkl)

Step 4 : Return Feature_set[]

Testing

Input: Test Dataset which contains various test instances TestDBLits [], Train dataset which is built by training phase TrainDBLits[] , Threshold Th.

Output: HashMap <class_label, SimilarityWeight> all instances which weight violates the threshold score.

Step 1: For each read each test instances using below equation

$$testFeature(m) = \sum_{m=1}^{n} (.\ featureSet[A[i] \dots\dots A[n] \leftarrow \text{TestDBLits})$$

Step 2: extract each feature as a hot vector or input neuron from $testFeature(m)$ using below equation.

Extracted_FeatureSetx [t……n] = $\sum_{x=1}^{n}(t) \leftarrow testFeature$ (m)

Extracted_FeatureSetx[t] contains the feature vector of respective domain

Step 3: For each read each train instances using below equation

$$trainFeature(m) = \sum_{m=1}^{n} (. featureSet[A[i] \dots\dots A[n] \leftarrow \text{TrainDBList})$$

Step 4: extract each feature as a hot vector or input neuron from $testFeature(m)$ using below equation.

Extracted_FeatureSetx [t……n] = $\sum_{x=1}^{n}(t) \leftarrow testFeature$ (m)

Extracted_FeatureSetx[t] contains the feature vector of respective domain.

Step 5: Now map each test feature set to all respective training feature set

$$weight = calcSim\ (\text{FeatureSetx} \,||\, \sum_{i=1}^{n} \text{FeatureSety[y]})$$

Step 6: Return Weight

RESULTS AND DISCUSSIONS

CONCLUSION

Recent research suggests that the best modeling of temporal sequences is still an unresolved subject that is being addressed using recurrent neural networks. Because of their capacity to preserve both short- and long-term context information in their cell architectures, CNN have been frequently utilized for modeling sequences, albeit it is unclear how to fully use this feature. Several authors, for example, have used numerous CNN layers to mimic various scales of context, with the goal of introducing constraints relating to larger speech structures such as connected phonemes, syllables, phrases, or sentences. CPUs will have at least 2GB of RAM and a faster CPU as technology improves. It lacks Python's numpy modules, but it's a fairly stable language with a vast supporting library that's ideal for extracting lexical emotions. Before we go, we'll use MySQL to store the classified data we've extracted from raw data, making keyword and emotional searches easy. For categorizing and coordinating, the latter CNN and DCNN are utilized.

REFERENCES

[1] Makino, Takaki, et al. "Recurrent neural network transducer for audio-visual speech recognition." 2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019.

[2] Lee, Yong-Hyeok, et al. "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model." Applied Sciences 10.20 (2020): 7263.

[3] Isobe, Shinnosuke, et al. "Efficient Multi-angle Audio-visual Speech Recognition using Parallel WaveGAN based Scene Classifier." (2022).

[4] Bekmanova, Gulmira, et al. "Emotional Speech Recognition Method Based on Word Transcription." Sensors 22.5 (2022): 1937.

[5] Kwon, Soonil. "A CNN-assisted enhanced audio signal processing for speech emotion recognition." Sensors 20.1 (2019): 183.

[6] Jeon, Sanghun, and Mun Sang Kim. "End-to-End Lip-Reading Open Cloud-Based Speech Architecture." Sensors 22.8 (2022): 2938.

[7] Afouras, Triantafyllos, et al. "Deep audio-visual speech recognition." IEEE transactions on pattern analysis and machine intelligence (2018).

[8] Zhou, Pan, et al. "Modality attention for end-to-end audio-visual speech recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[9] Su, Rongfeng, et al. "Cross-domain deep visual feature generation for mandarin audio–visual speech recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019): 185-197.

[10] Ramadan, Rabie A. "Detecting adversarial attacks on audio-visual speech recognition using deep learning method." International Journal of Speech Technology (2021): 1-7.