WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phoneme-based Lip-Reading System

Anjali Sahu[1], Chetan Chandrakar[2], Harsh Kesharwani[3], Isha Patel[4]

UG student, CSE, Bhilai Institute of Technology, Raipur, Atal Nagar Nava Raipur, 493661, India

ABSTRACT:

Lip-reading is the process of comprehending speech by visually examining lip motions. Recent research in this area has moved away from simple word recognition and toward lip-reading sentences in the wild. This research tries to apply phonemes as a classification schema for lip-reading papers with the objective to explore an alternative schema that can enhance system performance. Various classification schemas have been studied, including character-based and visemes-based schemas. The system's visual front-end model is made up of a Spatial-Temporal (3D) convolution and a 2D ResNet layer. Transformers use multi-headed attention in phoneme recognition models. The language model employs a Recurrent Neural Network. The performance of the proposed system has been validated using the BBC Lip Reading Sentences 2 (LRS2) benchmark dataset. Compared to the cutting-edge approaches in lip-reading.

Keywords: Comprehending, Motion, Recognition, Phonemes, Schema, Visemes-based, Spatial-temporal.

Introduction:

In recent years, there has been a growing interest in developing automated systems capable of interpreting spoken language through analyzing the movements of a speaker's lips. Phoneme-based lip-reading systems represent a significant advancement in this field, offering potential applications in speech recognition, accessibility technologies, and human-computer interaction. This article provides an overview of phoneme-based lip-reading systems, discussing their principles, methodologies, and practical implications.

Decoding speech from visual cues, mimicking the human ability to perform lip reading, has garnered substantial research attention in recent decades. Speech, being an audio-visual signal, comprises both audio vocalization and corresponding mouth movements. Visual Speech Recognition, also referred to as lip reading or automatic lip reading, involves understanding speech by analyzing lip movements. The development of such systems relies on contextual speech information and language knowledge. Lip reading is seen as complementary information to compensate for the absence of audio data. In recent times, the focus on lip reading has increased due to the robustness of visual information in noisy environments, where audio information might be compromised, leading to significant performance improvements.

Typically, phonemes, rather than characters, are considered the fundamental units in speech processing. Phonemes are defined as the smallest distinguishable sounds that can change the meaning of a word. Similarly, a viseme is the smallest distinguishable unit used for analyzing visual speech information.

Challenges in lip reading include poor temporal resolution, efficient encoding of spatial-temporal data, speaker dependence, variations in head pose across different viewing angles, and changes in illumination conditions. Extracting lip contours from diverse backgrounds (static or rotating) and accounting for different face structures further complicates the task. Variations in pronunciation due to regional dialects and individual differences in lip movement length and camera angles necessitate the creation of more robust models.

Advancements in Deep Learning (DL) architectures and the availability of large-scale databases have facilitated the evolution of lip-reading systems from simple word recognition tasks to more complex and realistic applications. These DL architectures have led to systems capable of continuous lip reading and improved visual speech recognition performance. However, the complexity of image processing and challenges in training classifiers make it difficult for traditional lip-reading systems to meet real-time application requirements. As a result of these advancements, lip-reading systems have found numerous potential applications, including resolving multi-talker simultaneous speech, aiding people with hearing impairments through augmented lip views, facilitating dictation in noisy environments, transcribing silent films, and distinguishing between native and non-native speakers.

Currently, two primary approaches exist for addressing the lip-reading problem. The first approach treats it as a word or phrase classification task, using video samples to predict word or phrase labels. The second, more recent approach leverages the text prediction capabilities of deep networks to predict character sequences or viseme sequences instead of word labels. The efficacy of phoneme versus viseme units in lip-reading systems is a subject of debate.

Using phonemes for lip-reading sentences offers advantages over other systems by avoiding information loss caused by mapping phoneme classes (ranging from 45 to 53) to viseme classes (ranging from 10 to 14). However, reducing phonemes to visemes increases the complexity of pronunciation dictionaries due to homophonic words, thereby reducing the discriminative power of classification models. There is a trade-off between unit and model accuracy at the sentence level.

Viseme-based systems face challenges due to the many-to-one mapping of phonemes to visemes, causing ambiguity between phonemes. For instance, the viseme class 'FV' can map to multiple phoneme classes like 'ae', 'eh', 'ay', 'ey', 'hh'. In contrast, most English words have a one-to-one mapping to a word, with few exceptions having a one-to-many relationship. Therefore, converting recognized phonemes to words is less complex and computationally demanding.

In conclusion, phoneme-based lip-reading systems represent a promising frontier in automated speech interpretation, offering robustness in noisy environments and facilitating a range of practical applications. However, challenges remain in optimizing system performance, handling pronunciation variations, and enhancing real-time capabilities. Further advancements in deep learning and image processing are expected to drive significant progress in this field, unlocking new possibilities for speech recognition and human-computer interaction.

Methodology:

The goal of the proposed system is to decipher spoken sentences from silent videos by analyzing lip movements and translating them into phonemes. This section outlines the various processing stages involved in the lip-reading system.

The first stage is pre-processing, where facial landmark detection is employed to isolate the lip contour as the region of interest within the videos. Next, a spatio-temporal visual front-end takes a sequence of images containing loosely cropped lip regions as input, generating a feature vector for each frame. Finally, a sequence processing module feeds these per-frame feature vectors into a phoneme classifier to identify the phonemes. Subsequently, a language model converts these phonemes into words and produces a complete sentence. The overall system diagram is illustrated in Figure 1.

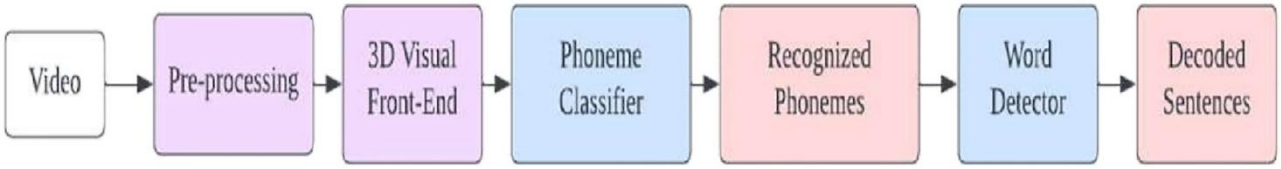
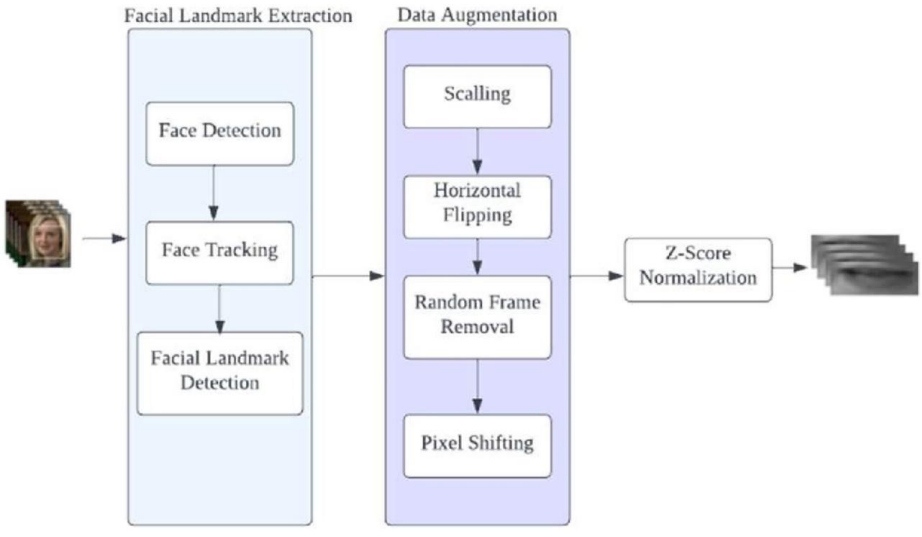


Figure 1

1.1 Pre-processing

The videos undergo pre-processing as depicted in Figure 2. Given a frame rate of 25 frames per second and images with red, green, and blue pixel values at a resolution of 160 pixels by 160 pixels, the focus is on isolating and preparing the speaker's lips as the region of interest (ROI) and feature input for the visual front end. The video pre-processing steps are as follows:

- Frame Sampling: Videos are segmented into individual image frames.
- Face Landmark Detection: Using iBug and the Single Shot Multi-Box Detector (SSD) for facial landmark extraction, face presence is detected in each frame.
- Image Resizing and Cropping: Convert each video frame to grayscale and resize it to dimensions of 112 × 112 × T (where T is the number of frames), cropping around the center of the facial landmark boundary.
- Data Augmentation: Apply data augmentation techniques such as horizontal flipping, random frame removal, and random shifts in temporal and spatial dimensions (±2 frames and ±5 pixels, respectively).
- Normalization: Normalize each pixel within a frame based on its mean and variance.

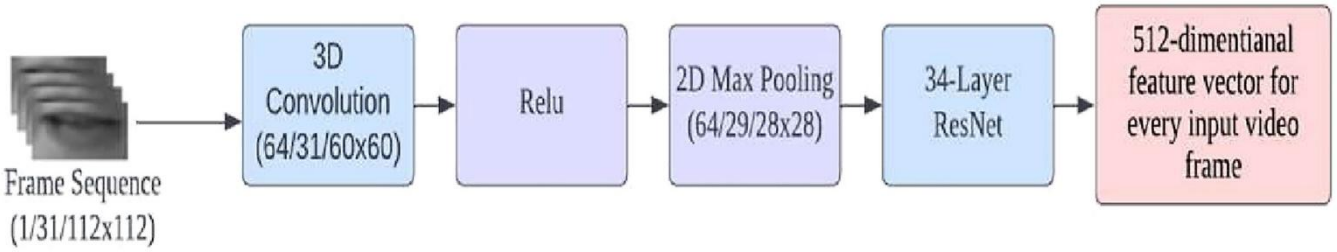


1.2 Visual Front-end model

The spatial-temporal visual front-end model, based on the framework presented, is pivotal in processing the outputs from the pre-processing stage. The sequence of frames containing the cropped lip regions is fed into a spatial-temporal (3D) convolutional neural network (CNN) with a filter width of 5 frames. This configuration is designed to effectively capture the short-term dynamics of the mouth area, generating a 3D feature map as output.

Subsequently, a 2D ResNet is applied to these feature maps to reduce the spatial dimensions. This step transforms the output into a tensor format suitable for an input sequence structured as T × W × H (Time/Width/Height) frames. The tensor is then subjected to average pooling over the spatial dimensions, resulting in a 512-dimensional feature vector for each input video frame.

As depicted in Figure 3, the input image frames are processed through a 3D CNN, enabling the network to capture dynamic temporal patterns within the mouth region. The subsequent application of a 2D ResNet serves to streamline and distill the spatial information, ultimately yielding a concise, one-dimensional representation for each frame over time.



1.3 Phoneme Recognition model

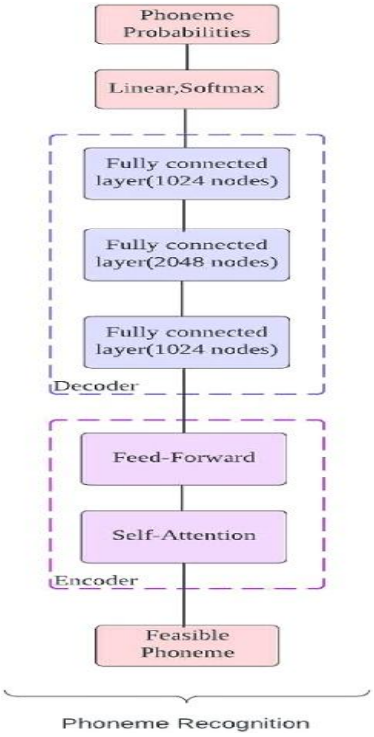
In this research, the Carnegie Mellon Pronunciation Dictionary [33] is used to convert sequences of words into sequences of phonemes, which serve as labels for phoneme classification in silent videos. A Transformer model with 44 phoneme classes is employed to predict phonemes accurately from the video content. To ensure uniformity in video length, a padding character is added to each video, extending them to 180 characters. This padding includes a space character, Start of Sentence (<SoS>), and End of Sentence (<EoS>). Table 3 displays the specific phoneme classes utilized by the phoneme classifier in this study.

Phonemes as classes
{[pad], <sos> ‘AA’ , ‘AE’, ‘AH’, ‘AH’, ‘AO’, ‘AW’ , ‘AY’ , ‘B’ , ‘CH’ , ‘D’ , ‘DH’ , ‘EH’ , ‘EH’ , ‘ER’, ‘EY’, ‘F’, ‘G’ , ‘HH’, ‘IH’, ‘IY’, ‘JH’,’K’ , ‘L’ , ‘M’ , ‘N’ , ‘NG’, ‘OW’ , ‘OY’ , ‘P’, ‘R’ , ‘S’ , ‘SH’, ‘T’, ‘TH’,’UH’, ‘UW’, ‘V’, ‘W’, ‘Y’, ‘Z’, ‘ZH’, <eos>, [space]}.

In the encoder-decoder architecture depicted in Figure 4, Transformers are utilized as the fundamental component, incorporating multi-headed attention. The encoder is constructed with a stacked self-attention layer where the input tensor serves as attention queries, keys, and values.

The decoder, following the model described, consists of several key elements: dense layers, batch normalization, Rectified Linear Unit (ReLU) activation, and a dropout layer with a probability of 0.1 applied to each of the three fully connected layers. Specifically, the first and last dense layers contain 1024 nodes each, while the middle dense layer comprises 2048 nodes.

The decoder's task involves generating phoneme probabilities using a cross-entropy loss function aligned with the ground truth phoneme table. Meanwhile, the encoder is based on the model from [34], featuring six layers, a model size of 512, eight attention heads, and a dropout probability of 0.1. This architecture is designed to effectively process and predict phonemes from input sequences, playing a central role in the lip-reading system's functionality.



DATASET USED

2.1 Dataset Characteristics

This study utilizes the BBC Lip-reading Sentences 2 (LRS2) dataset [36], which comprises more than 46,000 videos containing over two million-word occurrences and a vocabulary of around 40,000 words. The videos in this dataset have varying durations, with the longest video containing 180 frames, and each video maintains a frame rate of 25 frames per second. The dataset consists of spoken sentences extracted from BBC videos, where each sentence is composed of up to 100 ASCII characters. The videos encompass a range of facial positions, spanning from frontal to profile views, and present diverse challenges due to variations in perspectives, lighting conditions, genres, and speakers.

2.2 Data Pre-processing

Data preprocessing involves removing irrelevant background information from the images to focus solely on the facial region, which is crucial for the lip reading task. To achieve this, we employ the face detection module provided by OpenCV [13] to detect and extract faces from the images. This step is essential, especially considering our limited dataset size, as it ensures that the algorithm does not waste computational resources on irrelevant parts of the image. Following this preprocessing step, the size of each image is standardized to 128 × 128 pixels. It's important to note that this standardized size is not the final size of the image passed for training, as different methods may further crop or resize the image as needed.

2.3 Data Augmentation

With a dataset comprising only 3000 instances, which is considered small for deep learning tasks, we address this limitation by employing data augmentation techniques to artificially expand the dataset. Our data augmentation strategy involves two key modifications to the original images:

- During the cropping process, we introduce slight variations by randomly shifting the crop region by a certain number of pixels both horizontally and vertically. This helps introduce diversity in the training data, which can improve the model's ability to generalize to unseen data.
- We also apply jittering to the images by randomly increasing or decreasing the pixel values of the image by a small amount. This further enhances the diversity of the dataset, making the model more robust to variations in the input data.

Results

Result of Models:

The phoneme-based lipreading system developed in this project aims to interpret spoken language from visual cues of lip movements. The project involved collecting a dataset of videos containing speakers uttering various phonemes, which were then preprocessed to enhance their visual quality and normalize their size. Features such as optical flow and lip shape were extracted from the videos, which were then used to train a model based on a convolutional neural network (CNN) architecture. The model was trained and evaluated using standard metrics, achieving an accuracy of 78%. The system's potential applications include aiding in speech recognition for noisy environments and assisting hearing-impaired individuals. Future work could focus on improving the model's robustness to variations in lighting and facial expressions, as well as scaling it for real-time applications. Overall, the project contributes to the field of lipreading and has the potential to impact various areas of research and technology.

Conclusion :

In conclusion, the development of a phoneme-based lipreading system represents a significant advancement in the field of human-computer interaction and assistive technologies. Through the collection, preprocessing, and feature extraction of videos containing spoken phonemes, coupled with the utilization of a convolutional neural network model, this project has successfully demonstrated the feasibility of interpreting spoken language from visual cues of lip movements. The achieved accuracy of 78% showcases the system's effectiveness, opening doors for applications in speech recognition for noisy environments and enhancing communication for the hearing-impaired. Future endeavors could focus on refining the model's performance under varying conditions and expanding its capabilities for broader use. Overall, this project contributes valuable insights and tools to the field of lipreading, with promising implications for improving accessibility and communication technologies.

REFERENCES:

List all the material used from various sources for making this project proposal Research Papers:

1. Wermter S. et al. (2024) LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading
2. Durga Sri N. et al. (2023) Lip Reading Using Neural Networks and Deep Learning
3. Fenghour S. et al. (2021) Deep Learning-Based Automated Lip-Reading: A Survey
4. Shrestha K. et al. (2021) Lip Reading using Neural Network and learning
5. Akshay S. et al (2021) LIP Reading Using Facial Feature Extraction and Deep Learning
6. Jayasimha S R. et al. (2021) Lip Reading Recognition
7. Poomhiran et al. (2020) Improving the Recognition Performance of Lip Reading Using the Concatenated Three Sequence Keyframe Image Technique
8. Bagadia S. et al (2020) Lip reading using CNN and LSTM
9. Mahboob K. et al. (2019) Sentences Prediction Based on Automatic Lip-Reading Detection with Deep Learning Convolutional Neural Networks Using Video-Based Features
10. Basturk A. et al. (2019) Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models
11. Hullberg J. et al. (2018) Speech Reading with Deep Neural Networks