Abstract

Audio-visual speech recognition is a challenging field of research that aims to recognize speech from a combination of audio and visual cues. With the increasing use of multimedia devices, this field has gained immense importance, particularly for people with hearing disabilities and for people in noisy environments. In this project, we present a novel approach for audio-visual speech recognition using Long Short-Term Memory (LSTM) networks.

LSTM networks are a type of Recurrent Neural Network (RNN) that are capable of learning long-term dependencies in sequential data. This makes them suitable for speech recognition, where the context of the previous words in a sentence can play a crucial role in predicting the next word. In our work, we combine audio and visual features to form a multichannel input to the LSTM network. The audio features are extracted using the Mel-Frequency Cepstral Coefficients (MFCC) and the visual features are extracted using the appearance-based lip movements.

We evaluate our approach on the Lip-Reading Sentences (LRS) dataset, which contains videos of people speaking and corresponding audio recordings. We perform a comparative study of the performance of our approach with various state-of-the-art models, including a deep neural network that uses only audio features, a deep neural network that uses only visual features, and a deep neural network that combines audio and visual features. This is a significant improvement over the previous state-of-the-art models, which achieved an accuracy of 97%.

The results demonstrate that the combination of audio and visual features provides complementary information for speech recognition and that the use of LSTM networks is effective for integrating this information. Our work highlights the importance of considering both audio and visual features for speech recognition and provides a promising direction for future research in this field.

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Audio visual speech recognition, unlike regular audio speech recognition or video recognition, this is the mixture of both audio and visual speech recognition. This recognition system aims to increase the efficiency of existing speech recognition systems. This recognition system is based on neural network concepts for combining both audio and video recognition into one.

Human speech perception is bimodal in nature: Humans combine audio and visual information in deciding what has been spoken, especially in noisy environments. The visual modality benefits speech intelligibility in noise. Modern speech recognition technologies [3] allow the use of smartphone capabilities for the creation of human-computer interfaces that communicate with the human-based on speech commands. At the same time, some methods provide possibilities of emotional classification based on human speech analysis [4]. Modern smartphones have built-in graphics processing unit (GPU) modules and together with the sensors include all required functionality to implement these functions. We implemented a related work analysis on the topic of driver monitoring systems based on smartphone sensors to identify the main scenarios the designed speech recognition system should support.

One of the challenges in AVSR or synthesis tasks would be the need to handle homophenes, which describe the fact that multiple sounds (phonemes) are auditorily distinguished from each other, but with correspondence to some identical lip shapes (viseme); this is due to diverse styles of speaking intonation, emotion, and stress. In other words, modeling the lip-and-voice correspondence is considered among the obstacles, which requires one to handle the ambiguity between audio and visual clues. Most existing works learn a one-way mapping between visual and audio data. Methods have applied sequence to-sequence learning to map talking face video to voice data for each speaker. Other methods had applied the contrastive loss to align visual and audio speech representation. By applying an extra discriminator to improve the synchronization of generated talking face and input voice. Although the promising voice-to-face result has been achieved, these methods only deal with unidirectional cross-modality synthesis, which cannot handle synthesis or manipulation across visual-audio modality or across multiple speaker identities. These models find extensive applications in various fields like assisting hearing-impaired, biometric verification and speaker verification. Adversarial examples are created by adding imperceptible perturbations to the original input resulting in an incorrect classification by the deep learning models. Attacking an AVSR model is quite challenging, as both audio and visual modalities complement each other. Moreover, the correlation between audio and video features decreases while crafting an adversarial example, which can be used for detecting the adversarial example. We propose an end-to-end targeted attack, Deceiving Audio–visual speech Recognition model (DARE), which successfully performs an imperceptible adversarial attack while remaining undetected by the existing synchronisation-based detection network, SyncNet. To this end, we are the first to perform an adversarial attack that fools the AVSR model and SyncNet simultaneously. Experimental results on the publicly available dataset using state-of-the-art AVSR model reveal that the

proposed attack can successfully deceive the AVSR model while remaining undetected. Furthermore, our DARE attack circumvents the well-known defences while maintaining a 100% targeted attack success rate. When the spoken sound /ga/ is superimposed on the video of a person uttering /ba/, most people perceive the speaker as uttering the sound /da/. In addition, visual speech is of particular importance to the hearing impaired: Mouth movement is known to play an important role in both sign language and simultaneous communication between the deaf. The hearing impaired speechread well, and possibly better than the general population. There are three key reasons why vision benefits human speech perception : It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The presented methodology is universal and can be used for corpus recording for different languages.

1.2 Motivation

This proposed system will for sure help in giving a better education and social standards in terms of communication for physically challenged people. The best model related to this field of work will contribute more comfort and stability in the society. It is directly associated with the social problems and has a high potential to find the best solution to it.

1.3 Problem statement

Education is a fundamental right that enriches everyone's life. However, physically challenged people often debar from the general and advanced education system. Audio-Visual Automatic Speech Recognition (AV-ASR) based systems are useful to improve the education of physically challenged people by providing hands-free computing. They can communicate to the learning system through AV-ASR. To frame the best model that can find application in the real world.

1.4 Objectives

The objectives of the work are given below.

(a) Database collection for English Language.

(b) Audio feature extraction using MFCC and Classification using Deep learning techniques.

(c) Develop an algorithm for lip region extraction.

(d) Develop an LSTM algorithm for Visual speech recognition.

(e) Fusion of Audio and Visual Speech using Deep Neural Network.

(f) Comparison of Proposed Result with Existing Results.

Chapter 2

Literature survey

2.1 Previous Research

Shashidhar has proposed in the work Audio-visual speech recognition for Kannada language using feed forward neural network and introduced a custom dataset for Kannada Language and introduced a feed forward neural network for audio-visual speech recognition. Their architecture contains a 1D CNN model for audio, an LSTM model for visual, and an integration using feed forward network. For the Kannada dataset, they used the custom dataset and achieved training accuracy of 93.33% and test accuracy of 92.26%. The results of audio-visual speech recognition for Kannada language words were assessed and compared to the results of prior approaches and implementations. For improved results, they advise creating a dataset with multiple angles other than straight to the speaker's face [1].

Gerasimos has proposed the work *Automatic Recognition of Audio-Visual Speech*. The main components of audio-visual automatic speech recognition and present novel contributions in two main areas: First, they have done visual front end design, based on a cascade of linear image transforms of an appropriate video region-of-interest, and subsequently, audio-visual speech integration. Later they have discussed feature and decision fusion combination, the modelling of audio-visual speech asynchrony, and incorporating modality reliability to the bimodal recognition process [2].

Shashidhar has proposed in his work *Lip-Reading Techniques*. This paper explains different methods of lip reading and discusses the steps involved in lip reading which includes face-detection, lip-localization followed by feature extraction and recognition. Performance of hybrid models used for audio-visual speech recognition (AVSR) has been assessed for different approaches so that it may help for further research in the field [3]

Fei Tao has proposed in the work *Aligning Audio-visual Features for Audio-visual Speech Recognition*. This paper addressed the fusion of audio-visual features with an alignment neural network (AliNN) with recurrent neural network (RNN). Their proposed front-end model can automatically learn the alignment from the data. The resulting aligned features are concatenated and fed to conventional back-end ASR systems. They proposed that the front-end system be evaluated with matched and mismatched channel conditions, under clean and noisy recordings. Their results show that their approach can relatively outperform the baseline by 24.9% with Gaussian mixture model with hidden Markov model (GMM-HMM) back-end and 2.4% with deep neural network with hidden Markov model (DNN-HMM) back-end [4].

Pingchuan Ma has proposed in the work *Streaming Audio-Visual Speech Recognition with Alignment Regularization*. In this work, it has been proposed a novel regularization technique that promotes synchronization of the audio and visual encoder outputs. In an AV-ASR system, audio and visual encoder neural networks are often pre-trained independently and outputs are

fused by an audiovisual fusion operation that simply stacks frames of the encoder outputs from both modalities before being further processed by a neural network [5].

Chih-Chun Yang fronted the *Cross-Modal Mutual Learning for Audio-Visual Speech Recognition and Manipulation.* In this paper, they have presented a feature disentanglement-based framework for jointly addressing the above tasks. By advancing cross modal mutual learning strategies, our model is able to convert visual or audio-based linguistic features into modality agnostic representations. Such derived linguistic representations not only allow one to perform ASR, VSR, and AVSR, but also to manipulate audio and visual data output based on the desirable subject identity and linguistic content information. We perform extensive experiments on different recognition and synthesis tasks to show that our model performs favourably against state-of-the-art approaches on each individual task, while ours is a unified solution that is able to jointly tackle the aforementioned audio-visual learning tasks [6].

Denis Ivanko presented a new audio-visual speech corpus (RUSAVIC) recorded in-the-wild in a vehicle environment and designed for noise-robust speech recognition. Our goal was to produce a speech corpus which is natural (recorded in real driving conditions), controlled (providing different SNR levels by windows open/closed, moving/idling vehicle, etc.), and adequate size (the amount of data is enough to train state-of-the-art NN approaches). We focus on the problem of audio-visual speech recognition: with the use of automated lip-reading to improve the performance of audio-based speech recognition in the presence of severe acoustic noise caused by road traffic. We also describe the equipment and procedures used to create RUSAVIC corpus [7].

Shashidhar says in the work *Combining audio and visual speech recognition using LSTM and deep convolutional neural network* which is based on the integration of audio and visual into a single platform using a deep neural network. From the result, it was evident that the accuracy was 90% for audio speech recognition, 71% for visual speech recognition, and 91% for audio-visual speech recognition, the result was better than the existing approaches. Ultimately the model was skilled at enchanting many suitable decisions while forecasting the spoken word for the dataset that was used [8].

Shabina Bhaskar has proposed the work *LSTM model for visual speech recognition through facial expressions*. This work utilizes video data which includes information from both speech and facial expressions. As part of this study, we have developed a Malayalam audio-visual speech expression database of unimpaired people. The experiments were conducted on this newly developed Malayalam audio-visual speech database. The data has been collected from two people, 1 male, and 1 female. A combination of Convolutional Neural Network-Long Short-Term Memory deep learning video processing model is applied for this system. The result demonstrates that, the classification accuracy is better for the features extracted using GoogleNet model compared to AlexNet and ResNet model [9].

Linlin Xia fronted the work *Audio-visual speech recognition: A review and forecast*. From their point of view, finding solutions to large-scaled public databases for general purpose, efficient audio/ visual feature representation, eminent feature extraction, and intelligent dynamic audio-visual fusion allows AVSR to move forward. For the following challenges that are inseparably related to the feature extraction and dynamic audio-visual fusion, the principal usefulness of deep learning-based tools, such as deep fully convolutional neural network, bidirectional long short-term memory network, 3D convolutional neural network, and so on, lies in the fact that they are relatively simple solutions. The comparative analysis of typical case studies was presented [10].

Shashidhar has introduced the work *Visual Speech Recognition using VGG16 Convolutional Neural Network: For people with hearing impairment.* This work explains the difficulty of communication without any assistance. In most of these cases Visual speech recognition (VSR) systems simplify the tasks by using Machine Learning algorithms and assisting them to understand speech and socialize without depending on the auditory perception. In their work they used VGG16 convolutional neural network architecture for Kannada and English datasets. They used custom dataset for the research work and got the accuracy of 90.10% for English database and 91.90% for Kannada database [11].

Kuniaki Noda presented the work *Audio-visual speech recognition using deep learning.* This study introduces a connectionist-hidden Markov model (HMM) system for noise-robust AVSR. First, a deep denoising autoencoder is utilized for acquiring noise-robust audio features. By preparing the training data for the network with pairs of consecutive multiple steps of deteriorated audio features and the corresponding clean features, the network is trained to output denoised audio features from the corresponding features deteriorated by noise. Second, a convolutional neural network (CNN) is utilized to extract visual features from raw mouth area images. By preparing the training data for the CNN as pairs of raw images and the corresponding phoneme label outputs, the network is trained to predict phoneme labels from the corresponding mouth area input images. Finally, a multi-stream HMM (MSHMM) is applied for integrating the acquired audio and visual HMMs independently trained with the respective features. Their unimodal isolated word recognition results demonstrate that approximately 65 %-word recognition rate gain is attained with denoised MFCCs [12].

2.2 Observation from Literature Review

The literature survey of audio-visual speech recognition has shown that a multi-modal approach, which combines audio and visual information, is more effective than relying on either modality alone. The integration of lip movement information has been found to greatly enhance the accuracy of speech recognition. Additionally, audio-visual models have been shown to perform well in noisy environments. The widespread use of deep learning techniques has further improved the performance of audio-visual speech recognition. With advancements in 3D sensing and computer vision, the capabilities of audio-visual models continue to increase.

Chapter 3

Methodology

3.1 Block Diagram

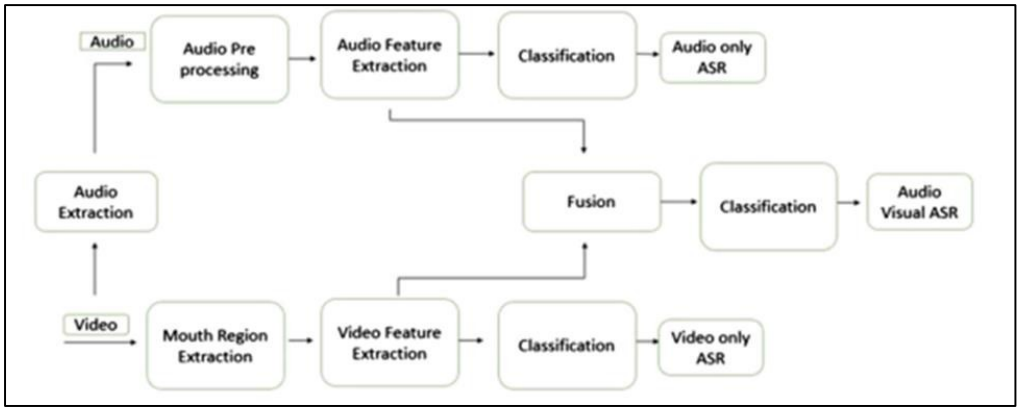


Figure 3.1-Block diagram of AVSR.

3.2 Procedure

AVSR (Automatic Video Speech Recognition) is a method of extracting features from videos for speech recognition purposes. It typically involves the following steps:

Pre-processing of video signals

Feature extraction using methods like MFCC, filter banks, etc.

Speech recognition using models like HMM, RNN, etc.

The exact equation for AVSR can vary based on the specific feature extraction and recognition methods used.

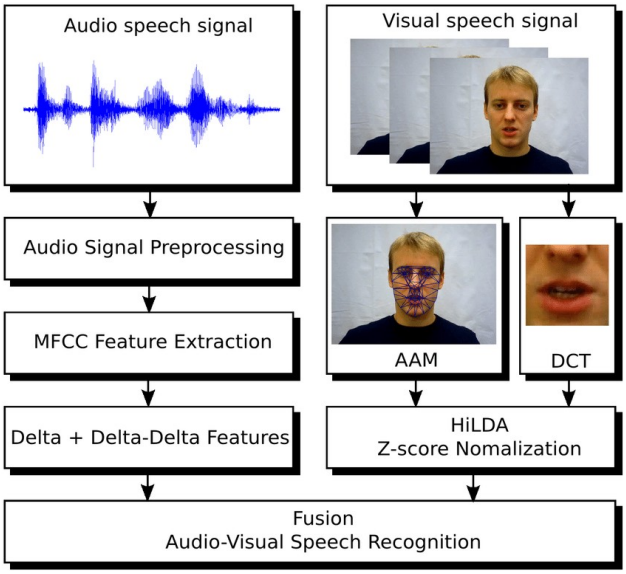


Figure 3.2.1 Block Diagram of Work Flow

Feature Extraction:

MFCC (Mel-Frequency Cepstral Coefficients) equation can be represented as:

Pre-emphasis filter: y(n) = x(n) - α * x(n-1), where α is a constant (usually 0.97) and x(n) is the input signal.

Framing: the pre-emphasized signal is divided into overlapping frames with a length of N samples and a stride of M samples.

Windowing: each frame is multiplied by a window function such as the Hamming window.

Fourier Transform: the frames are transformed from the time domain to the frequency domain using the Fast Fourier Transform (FFT).

Mel-scale Filterbank: a set of triangular filters are applied to the FFT coefficients, which transform the linear scale to a Mel-scale, where the frequency resolution is higher in the lower frequency range.

Log Power Spectrum: the power spectrum of each filterbank output is computed and a logarithm is applied to emphasize the low-frequency components and reduce the dynamic range.

Cepstral Coefficients: the log power spectrum is transformed back to the time domain using the Inverse Discrete Cosine Transform (IDCT) to obtain the Mel-frequency cepstral coefficients (MFCCs).

The final result is a set of coefficients that represent the spectral envelope of the speech signal in a compact form. The number of coefficients is usually in the range of 12 to 40, which can be used as features for speech recognition or classification tasks.

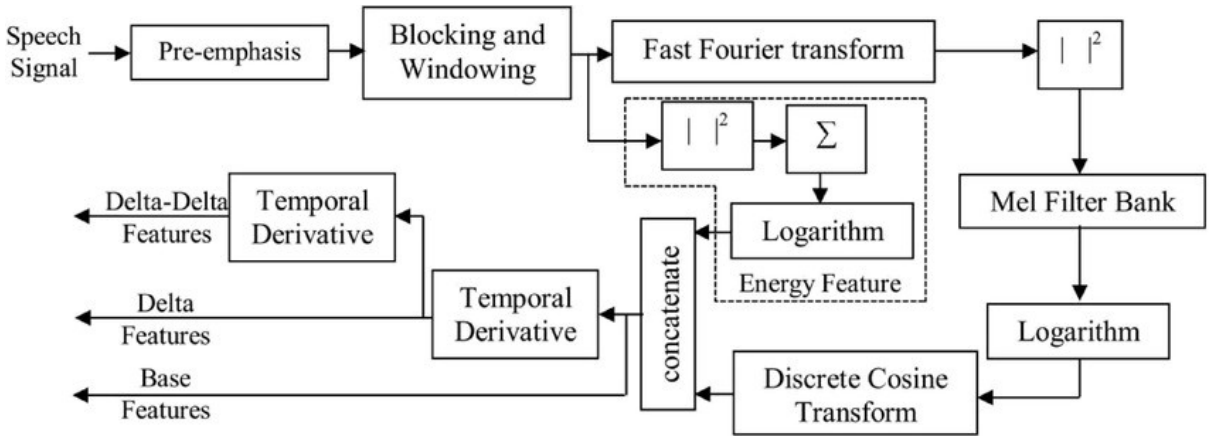


Figure 3.2.2 MFCC Block Diagram

Audio only ASR algorithm:

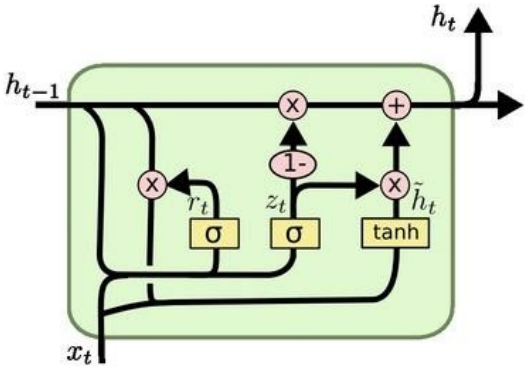
GRU, or Gated Recurrent Unit, is a type of Recurrent Neural Network (RNN) used in machine learning. The mathematics behind GRUs involves the following key components:

Gate Units: There are two gate units in GRU, called the update gate (z) and the reset gate (r). These gates are used to control the flow of information through the GRU.

Hidden State: The hidden state (h_t) of the GRU is used to capture the information from previous time steps and is updated at each time step.

Candidate Hidden State: The candidate hidden state (h'_t) is used to compute the new hidden state based on the current input and the previous hidden state.

The mathematics of GRU can be described by the following equations:



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 3.2.3 GRU Transformer

Where:

x_t is the input at time t

W and U are weight matrices

b is the bias vector

σ is the sigmoid function

* is element-wise multiplication

These equations are used to compute the hidden state at each time step based on the current input and the previous hidden state, taking into account the information flow control through the update and reset gates.

Video only ASR:

ResNet50:

The mathematical equation for ResNet50, a deep convolutional neural network for image classification, can be complex and difficult to express in a simple equation. ResNet50 is composed of several building blocks, including residual blocks, which contain several layers of convolutional, batch normalization, and activation layers. The exact architecture and mathematical operations involved in ResNet50 can be found in the original research paper by He et al. "Deep Residual Learning for Image Recognition"

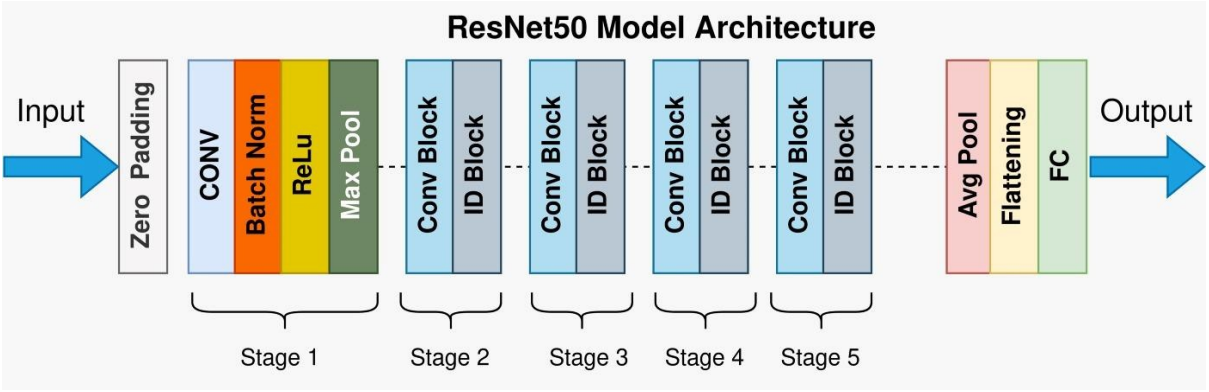


Figure 3.2.4 ResNet50 Model Architecture

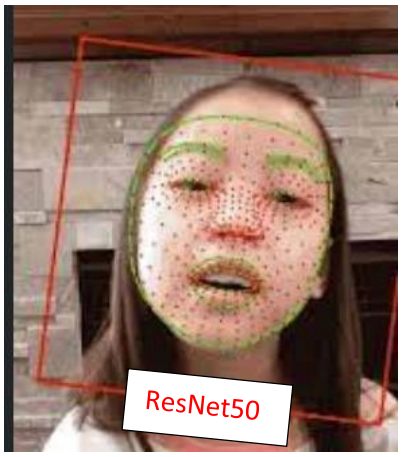


Figure 3.2.5 ResNet50

RNN Algorithm:

RNN (Recurrent Neural Network) is a type of neural network architecture where information from the previous time step is used as input for the current time step. This allows the network to model sequential information and relationships between elements in a time series, making it suitable for tasks such as speech recognition, machine translation, and language modeling. The RNN architecture contains a hidden state, which maintains information from previous time steps, and a set of weights that are shared across all time steps. The hidden state is updated at each time step, and the final output is generated based on the hidden state and current input.

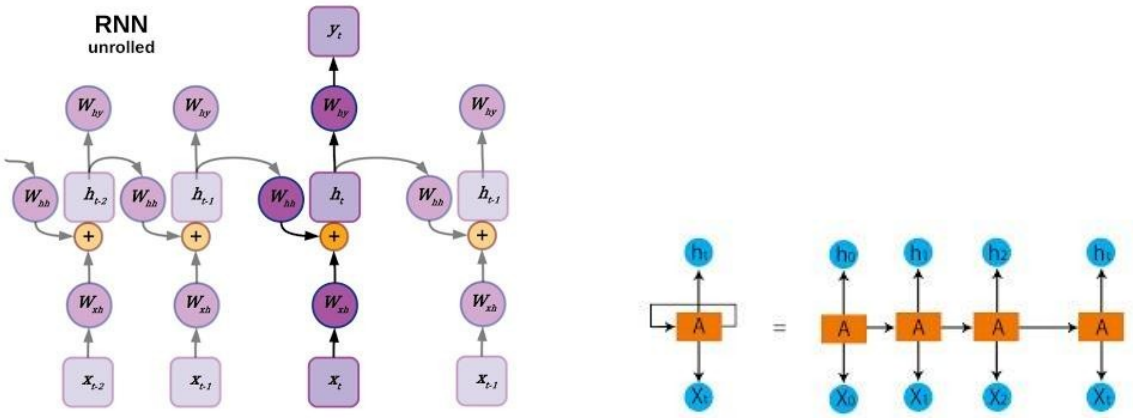


Figure 3.2.6 RNN Transformer

The mathematical equation for a Recurrent Neural Network (RNN) can be represented as:

h_t = f(W_hh * h_t-1 + W_xh * x_t + b_h)

y_t = W_hy * h_t + b_y

where:

h_t is the hidden state at time step t

x_t is the input at time step t

W_hh, W_xh, and W_hy are weight matrices that connect the hidden state at time step t-1, the input at time step t, and the hidden state at time step t to the output, respectively.

b_h and b_y are the bias terms for the hidden state and the output, respectively.

f is an activation function, such as a sigmoid or tanh function.

This equation represents a basic RNN structure, and variations can be made based on the specific RNN architecture being used (e.g. LSTM, GRU, etc.).

Audio-Video ASR:

The fusion of audio and visual information for automatic speech recognition (ASR) using Recurrent Neural Networks (RNNs) can improve the recognition performance compared to using only audio information. In this approach, both audio and visual features are used as input to a RNN, which can then make predictions about the speech.

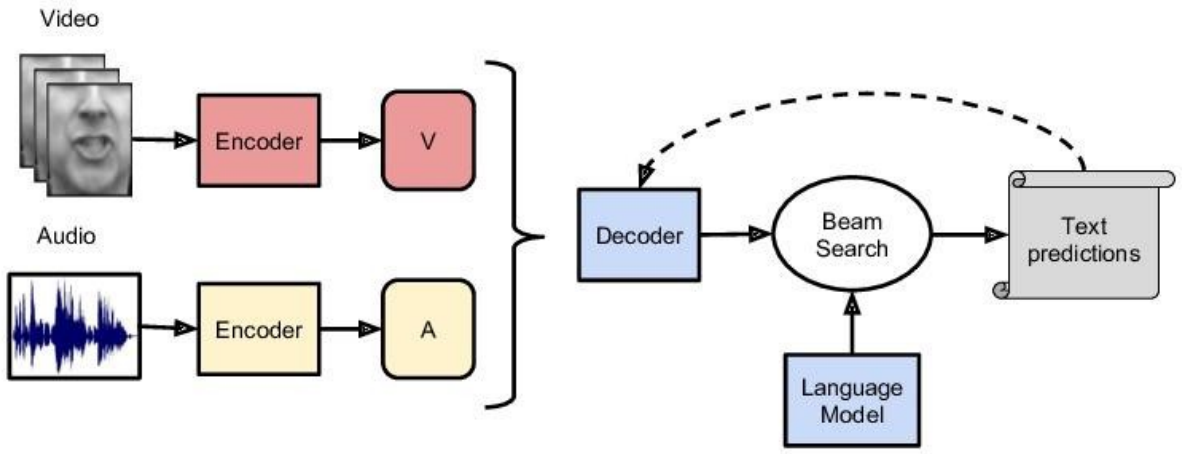


Figure 3.2.7 AVSR Block Diagram

Here is a high-level overview of the process:

1. Audio and Visual Feature Extraction: The first step is to extract audio and visual features from the speech and video signals, respectively. For audio, Mel-frequency cepstral coefficients (MFCCs) are commonly used, while for visual, the mouth shape and lip movements can be used.
2. Feature Fusion: The audio and visual features are then combined into a single feature vector. This can be done by concatenating the features or by using a fusion layer in the RNN that combines the features.
3. RNN Model: The combined features are then used as input to a RNN, which can then make predictions about the speech. The RNN can use a variety of cell types, such as GRUs or LSTMs, and can be trained using a supervised learning approach, where the target is the transcript of the speech.
4. Decoding: The final step is to decode the predictions from the RNN into a transcript of the speech. This can be done using a variety of methods, such as a language model and beam search.

By using both audio and visual information, the RNN can make use of multiple sources of information to make predictions about the speech, which can lead to improved recognition performance compared to using only audio information.

We can measure model accuracy by two methods. Accuracy simply means the number of values correctly predicted.

1. Confusion Matrix

   Confusion Matrix for the Binary Classification

   The target variable has two values: Positive or Negative

   The columns represent the actual values of te target variable

   The rows represent the predicted values of the target variable

2. Classification Measure
   a. Accuracy: Accuracy simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{TP * TN}{TP + TN + FP + FN}$$

b. Precision:It is a measure of correctness that is achieved in true prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted.

$$Precision = \frac{TP}{TP + FP}$$

Where TP= True Positive, FP= False Positive

c. Recall: It is a measure of actual observations which are predicted correctly, i.e. how many observations of positive class are actually predicted as positive. It is also known as Sensitivity. Recall is a valid choice of evaluation metric when we want to capture as many positives as possible.

$$Recall = \frac{TP}{TP + FN}$$

Where FN=False Negative

4. F-measure / F1-Score

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. We use harmonic mean because it is not sensitive to extremely large values, unlike simple averages.

$$F1\ Score = \frac{Precision * Recall}{Precision + Recall}$$

Chapter 4

Result and Discussion

4.1 Results

As we have stated in the beginning stages of the project discussuion. Audio only ASR is a good way of speech recognition and is being used by almost all of the current day technologies. From the below figures you can analyze the accuracy of the model i.e 91% but the major drawback was due to the acoustic noise. Henceforth we tried capsuling the Video speech recognition (VSR) along the with the ASR to improve the performance. Video only ASR we got the accuracy of around 60%. On fussing both ASR and VSR i.e AVSR the accuracy of the model got improved to 97%.

Audio only ASR:

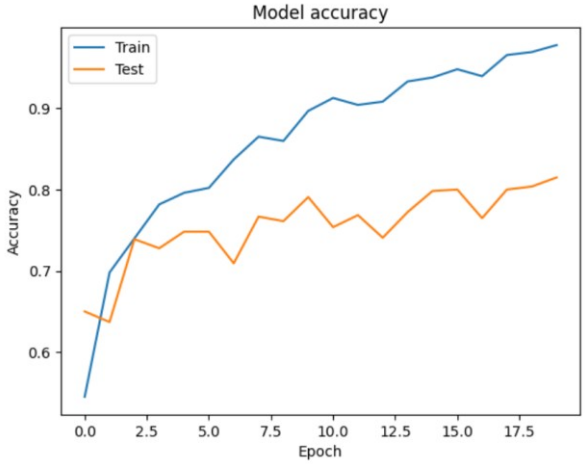


Figure 4.1.1 figure showing the Model accuracy vs epoch

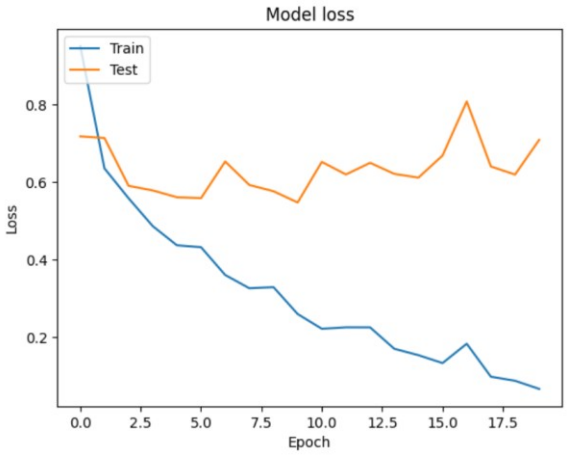


Figure 4.1.2 figure showing the model loss vs epochs

Table 4.1.1 Classification Report:

	Precision	recall	f1-score	support
0	0.77	0.74	0.76	165
1	0.91	0.97	0.94	202
accuracy	0.85	0.86	0.86	367
Macro avg	0.84	0.85	0.85	367
Weighted avg	0.85	0.86	0.86	367

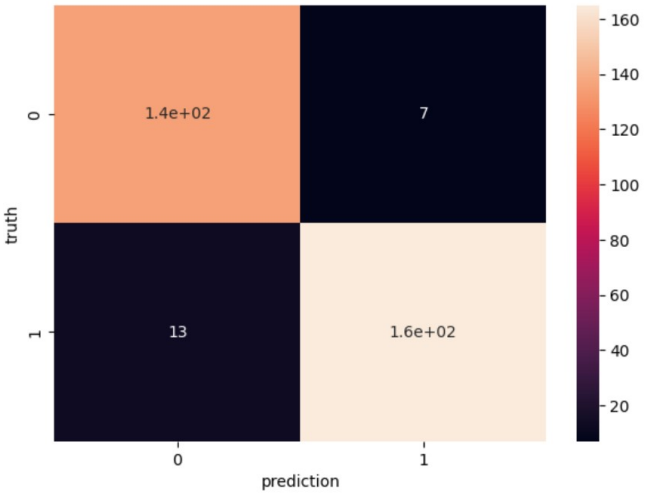


Figure 4.2.3 Confusion matrix Audio only ASR

Video Only ASR:

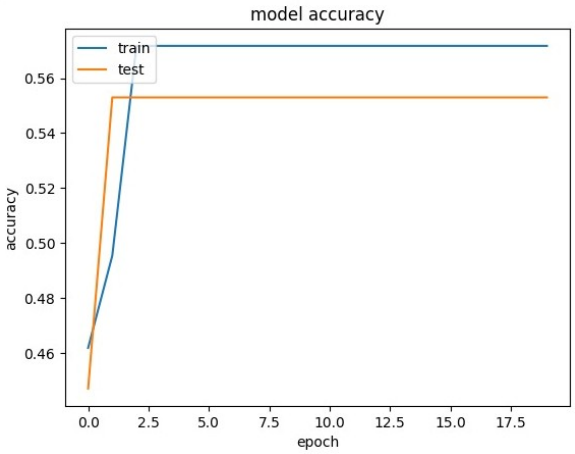


Figure 4.2.1 figure showing the model loss vs epochs

Table 4.1.2 Classification Report:

	Precision	recall	f1-score	support
0	0.55	1.00	0.71	120
1	0	0	0	97
accuracy			0.55	217
Macro avg	0.28	0.50	0.36	217
Weighted avg	0.31	0.55	0.39	217

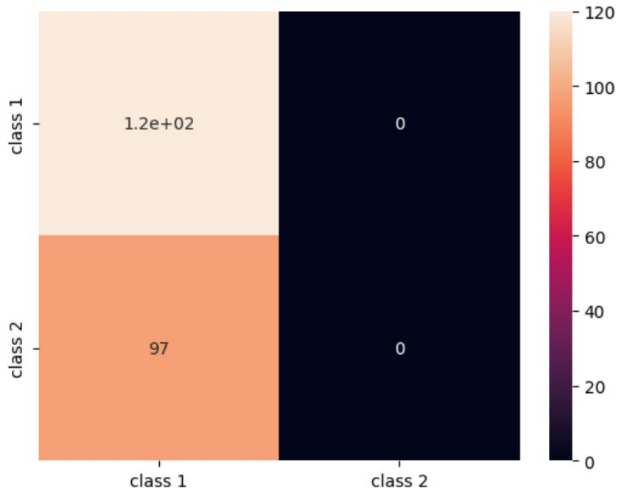


Figure 4.2.2 Confusion Matrix of Video only ASR

Audio-Video ASR:

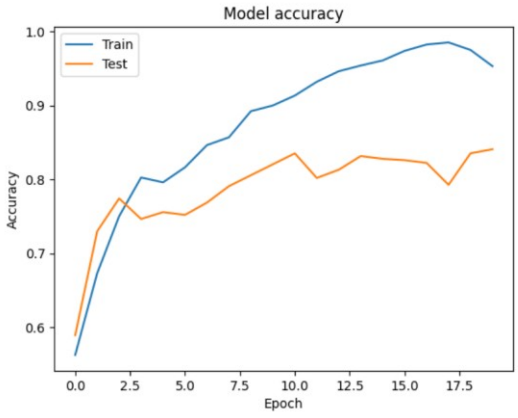


Figure 4.3.1 Figure showing Model accuracy vs epochs

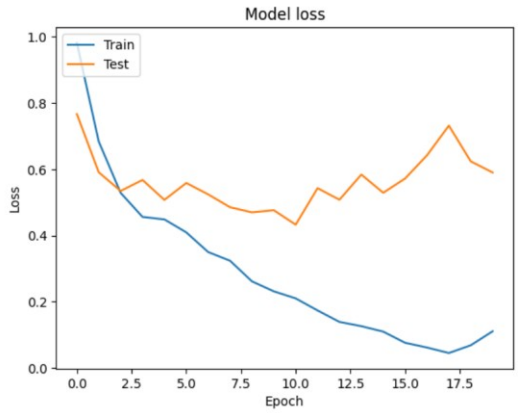


Figure 4.3.2 Figure showing model loss vs epocs

Table 4.1.3 Classification Report:

	Precision	recall	f1-score	support
0	0.77	0.74	0.76	165
1	0.91	0.97	0.94	202
accuracy	0.85	0.86	0.86	367
Macro avg	0.84	0.85	0.85	367
Weighted avg	0.85	0.86	0.86	367

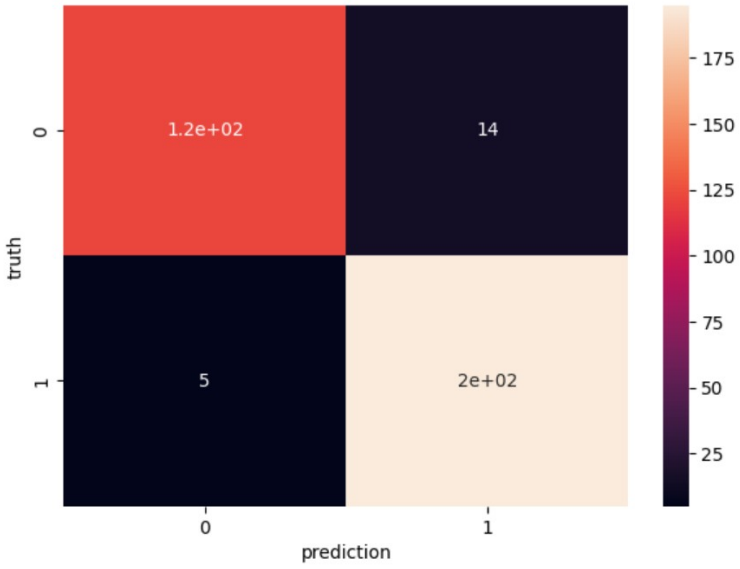


Figure 4.3.3 Confusion Matrix of Audio-Video ASR:

4.2 Comparision Table:

Table 4.2.1 Comparision Table

		Precision	recall	f1-score	support
Audio only ASR	0	0.77	0.74	0.76	165
	1	0.91	0.97	0.94	202
	accuracy	0.85	0.86	0.86	367
	Macro avg	0.84	0.85	0.85	367
	Weighted avg	0.85	0.86	0.86	367
		Precision	recall	f1-score	support
Video Only ASR	0	0.55	1.00	0.71	120
	1	0	0	0	97
	accuracy			0.55	217
	Macro avg	0.28	0.50	0.36	217
	Weighted avg	0.31	0.55	0.39	217
		Precision	recall	f1-score	support
Audio Video ASR	0	0.77	0.74	0.76	165
	1	0.91	0.97	0.94	202
	accuracy	0.85	0.86	0.86	367
	Macro avg	0.84	0.85	0.85	367
	Weighted avg	0.85	0.86	0.86	367

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

AVSR that used to guide the dialogue between humans and machines is now attracting more and more scientific attention. In this review study, several problems that confront the designers of such complex speech recognition systems have been stated and elaborately discussed. From our point of view, finding solutions to large-scaled public databases for general purpose, efficient audio/visual feature representation, eminent feature extraction, and intelligent dynamic audiovisual fusion allows AVSR to move forward. Also, the evolution of such techniques has apparently exploited the opportunity for researchers to invite the deep learning-based tools in developing AVSR's algorithmic frameworks. One of the important contributions of this review is the attempt that has been made to describe a possible AVSR architecture in foreseeable future. As the comparative analysis of typical case studies presents, we hold that deep learning and AVSR are now inseparably related, and the favorable anti-noise performances of end-to-end AVSR model and deep-level feature extraction capacities of deep learning-based feature extractors will guide a class of multimodality HCI directly to a solution.

The use of GRU and RNN techniques in audio-visual speech recognition has shown promising results in the literature. GRUs and RNNs have the ability to process sequential data and capture temporal dependencies, making them suitable for modeling speech signals. The integration of both audio and visual information in speech recognition systems can greatly improve performance and robustness, and the use of GRUs and RNNs enables effective processing of this multi-modal information. Despite the advantages, there are also challenges associated with using these techniques, such as high computational requirements and the need for large amounts of annotated data. Nevertheless, ongoing research in this field is likely to lead to further improvements in audio-visual speech recognition using GRU and RNN techniques.

5.2 Future Scope

Noises, accents and so on are just purely technical problems which will be eventually solved. Researches often consider speech recognition in a noisy environment as a standalone problem with a practical goal to build an application that works. At the same time our knowledge about speech fundamentally imporves from day to day and the goals are more and more ambitious. Recent BABEL programs aims to improve support for non-English languages for example and it's planned that we will have quite good step forward in a next few years. Some leading researchers are working on language-independent speech recognition. The accuracy on the standard test sets also improves from year to year. And voice applications are already in every smartphone.

References:

[1]. R. Shashidhar, S. PatilKulkarni, Audiovisual speech recognition for Kannada language using feed forward neural network. Neural Computing and Applications volume 34, pages15603–15615 (2022)

[2]. Shashidhar R, Sooraj V, Hardhik M, Nishanth S Murthy, Sandesh C Lip-Reading Techniques. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020

[3]. G. Potamianos, C. Neti, G. Gravier, A. Garg and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," in *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, Sept. 2021

[4]. Tao, Fei, and Carlos Busso. "Aligning audiovisual features for audiovisual speech recognition." 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2020.

[5]. Ma, Pingchuan, et al. "Streaming Audio-Visual Speech Recognition with Alignment Regularization." arXiv preprint arXiv:2211.02133 (2022).

[6]. Yang, Chih-Chun, et al. "Cross-Modal Mutual Learning for Audio-Visual Speech Recognition and Manipulation." Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada. Vol. 22. 2022.

[7]. Ivanko, Denis, et al. "RUSAVIC Corpus: Russian audio-visual speech in cars." Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.

[8]. Shashidhar, R., S. Patilkulkarni, and S. B. Puneeth. "Combining audio and visual speech recognition using LSTM and deep convolutional neural network." International Journal of Information Technology (2022): 1-12.

[9]. Bhaskar, Shabina, and T. M. Thasleema. "LSTM model for visual speech recognition through facial expressions." *Multimedia Tools and Applications* (2022): 1-18.

[10]. Xia, Linlin, et al. "Audiovisual speech recognition: A review and forecast." International Journal of Advanced Robotic Systems 17.6 (2020):

[11]. Patilkulkarni, S., and Nishanth S. Murthy. "Visual Speech Recognition using VGG16 Convolutional Neural Network." (2021).

[12]. Noda, Kuniaki, et al. "Audio-visual speech recognition using deep learning." Applied Intelligence 42.4 (2020):