Adversarial Attack on Audio-Visual Speech Recognition Model

MS (Research) Thesis

By

Saumya Mishra

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JUNE 2021

Adversarial Attack on Audio-Visual Speech Recognition Model

A THESIS

Submitted in fulfilment of the
requirements for the award of the degree
of

Master of Science (Research)

by

Saumya Mishra

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
JUNE 2021

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Adversarial Attack on Audio-Visual Speech Recognition Model** in the fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE (RESEARCH)** and submitted in the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2019 to June 2021 under the supervision of Dr. Puneet Gupta, Assistant Professor, Indian Institute of Technology Indore, Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

04/06/2021

Signature of the Student with date

(Saumya Mishra)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

4 June 2021

Signature of the Thesis Supervisor with date

(Dr. Puneet Gupta)

Saumya Mishra has successfully given her MS (Research) Oral Examination held on 24.09.2021

Signature of Chairperson (OEB)
Date: 24.09.2021

Signature of Thesis Supervisor
Date: 24.09.2021

Signature of Convener DPGC
Date: 24-09-2021

Somnath Dey

Signature of Head of Department
Date: 24/09/2021

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to people who, in one or the other way, contributed by making this time learnable, enjoyable, and bearable. At first, I would like to thank my supervisor **Dr. Puneet Gupta**, who was a constant source of inspiration during my work. With his constant guidance and research directions, this research work has been completed. His continuous support and encouragement has motivated me to remain streamlined in my research work. I am also grateful to **Dr. Somnath Dey**, HOD of Computer Science, for all his help and support.

I am thankful to **Dr. Somnath Dey** and **Dr. Vivek Kanhangad**, my research progress committee members, for taking out some valuable time to evaluate my progress all these years. Their valuable comments and suggestions helped me to improve my work at various stages.

My sincere acknowledgement and respect to **Prof. Neelesh Kumar Jain**, Director, Indian Institute of Technology Indore, for providing me the opportunity to explore my research capabilities at Indian Institute of Technology Indore.

I want to pay my gratitude to the respective authors for providing code and pretrained models. I am also grateful for Rob Cooper's help at BBC Research for providing the Lip Reading in the Wild (LRW) dataset.

I would like to express my heartfelt respect to my parents, my younger brother and my sister for providing their consistent support, love and care. I am thankful to all my friends and roommates for making this two years journey memorable.

Saumya Mishra

To my family and readers

ABSTRACT

The Audio-Visual Speech Recognition (AVSR) model is a favourable solution to predict text corresponding to the spoken words utilising both audio and face videos, specifically when the audio is corrupted by noise. These models have been widely used in applications like biometric verification, assisting hearing-impaired person, speaker verification in the multi-speaker scenario and event recognition in surveillance videos. However, these models are vulnerable to adversarial examples that can have profound implications such as distress to differently-abled and security breaches in surveillance systems. Adversarial examples are generated by adding imperceptible perturbations to clean samples with an intention to fool machine learning models. It is difficult to attack an AVSR model since audio and visual modalities complement each other. Furthermore, while generating an adversarial example, the correlation between audio and video features decreases, which can be used to detect the adversarial example for the AVSR model.

In this thesis, we introduce an end-to-end targeted attack, the Fooling Audio-visuaL Speech rEcognition, *FALSE*, that effectively performs an imperceptible adversarial attack while avoiding the detection by the existing synchronisation-based detection network (SyncNet). To the best of our knowledge, we are the first to perform an adversarial attack that simultaneously fools the AVSR model and SyncNet by introducing less distortion in audio and face videos. The experimental results show that the proposed attack successfully fool the state-of-the-art AVSR model on the publicly available dataset while avoiding the detection. Moreover, some well-known defences are easily circumvented by maintaining a 100% targeted attack success rate using our *FALSE* attack.

Keywords— Audio-Visual Speech Recognition; Cross-modality; Detection Network; Adversarial Attacks and Defenses.

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

AVSR	Audio-Visual Speech Recognition
FALSE	Fooling Audio-visuaL Speech rEcognition
ASR	Automatic Speech Recognition
LRW	Lip Reading in the Wild
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
BGRU	Bidirectional Gated Recurrent Unit
FGSM	Fast Gradient Sign Method
IGSM	Iterative Gradient Sign Method
MFCC	Mel-Frequency Cepstral Coefficient
BPDA	Backward Pass Differentiable Approximation
JPEG	Joint Photographic Experts Group
SNR	Signal to Noise Ratio
CNN	Convolutional Neural Network
FC	Fully Connected
ResNet	Residual Network
PFALSE	Fooling Audio-VisuaL Speech Recognition using Probabilities
RFALSE	Fooling Audio-VisuaL Speech Recognition by Restricting Video Distortions
AAVSR	Attacking only Audio-Visual Speech Recognition Model

Chapter 1

Introduction

Speech is an effective communication interface between humans and machine learning models. With the recent advancement of deep learning in several domains, Automatic Speech Recognition (ASR) models are proposed, which converts speech to the corresponding text, as shown in Fig. 1.1. These models are used in personal assistants like Google Assistant, Apple's Siri, Amazon's Alexa, Microsoft's Cortana and home electronic devices. However, the efficiency of these models decreases in the presence of noise. This can be overcome by

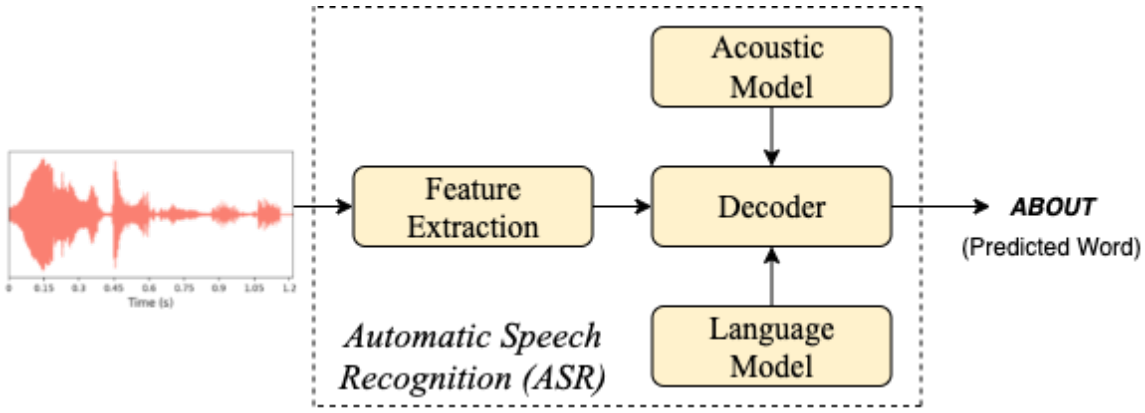

Figure 1.1: Illustration of Automatic Speech Recognition (ASR), ASR takes audio waveform as input and predicts the text corresponding to the given waveform.

either using speech enhancement techniques to remove noise [1] or adding visual features to speech [2]. Research has been done to show that there is a better sense of understanding when facial expression and lip movements derived from face videos are added to speech [3].

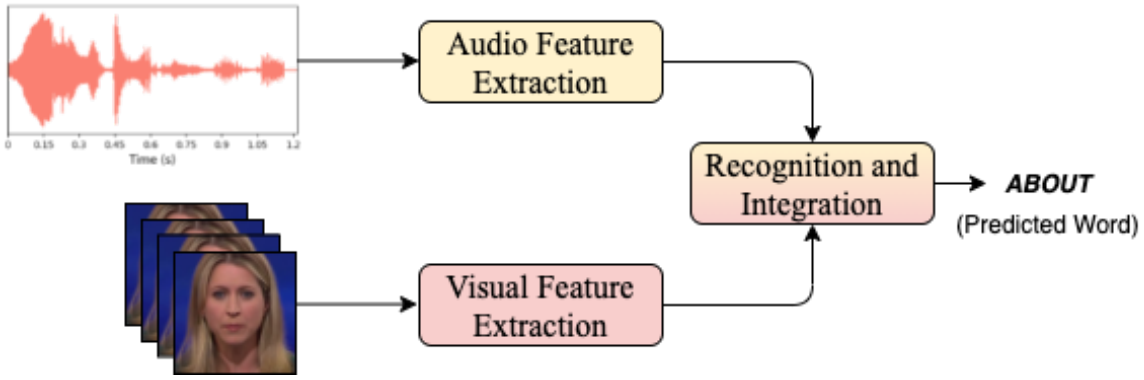

Figure 1.2: Overview of an Audio-Visual Speech Recognition (AVSR), AVSR takes audio waveform and face videos as input and predicts the word corresponding to the given input.

These characteristics increases the research in the area of Audio-Visual Speech Recognition (AVSR) [4]. The AVSR model extracts features from both audio and face videos to predict the text corresponding to the spoken word as shown in Fig. 1.2. The AVSR models are used in numerous real world applications such as: i) audio-visual speech separation [5]; ii) performing speech recognition even if one of the modalities (that is, either visual or audio) is noisy [6]; iii) biometrics verification [7]; iv) aiding hearing-impaired persons by providing transcriptions [8]; v) event recognition in surveillance videos [9]; vi) speaker verification in multi-speaker scenarios [10]; and vii) talking face synthesis [11].

The rest of this chapter is organized as follows. The motivation behind our work is elaborated in Section 1.1. In Section 1.2, we present the challenges of performing an adversarial attack on the AVSR model. The summary of thesis contributions is described in Section 1.3.

1.1 Motivation

With the advent of adversarial learning, adversarial attacks on image, audio and text have been extensively studied [12]. Results show that even the state-of-the-art deep learning models can be attacked by adding small perturbations (noise) to the original sample, producing erroneous classification results. However, the impact of adversarial attacks on the multimodal domain, specifically on AVSR models, are less explored. The AVSR models vul-

nerability to adversarial examples affects the efficacy of the applications outlined above. In this thesis, we investigated the AVSR model and analysed adversarial attacks against them. Hence, the research in this paper is motivated by the following questions: (i) Is it possible to perform an adversarial attack on the AVSR model whilst remaining undetected by the detection network? (ii) Is it feasible to generate the targeted adversarial samples that are difficult to be noticed by ordinary users? This thesis will cover in detail how our proposed attack is designed to address the above questions.

1.2 Challenges

In comparison with the existing image classification network, the AVSR models are more resistant to adversarial attacks. As compared to images, the temporal information is present in AVSR that acts as a defence method to mitigate the adversarial attacks [13]. Similarly, for fooling the AVSR model, we cannot use the existing adversarial attacks on video recognition models. This is because the region-of-interest is smaller in AVSR models, which leads to perceivable distortions. Moreover, generating an adversarial example for the AVSR model is more difficult than the existing ASR models as the AVSR works on two modalities that complement each other. That is to say, when we perform an attack on a single modality, the other modality attempts to reverse the prediction to the correct label. Therefore, it is challenging to perform the adversarial attack on the AVSR model compared to image classification, video classification and ASR model, which works on a single modality.

Generation of adversarial examples is usually done by backpropagating the gradients [14]. Due to some non-differentiable layers in the AVSR model, gradient backpropagation is not possible, thus preventing adversarial attacks. Even though if the adversarial examples are generated to fool the AVSR model, they can be easily detected [15] by the existing detection network SyncNet [10]. The detection network is devised on the idea that when perturbations are added to the original audio and face videos, the correlation between them decreases. That is, the correlation between adversarial audio and face videos is lower than the correlation between original audio and face videos.

1.3 Thesis Contribution

In this thesis, a novel adversarial attack on the AVSR model Fooling Audio-visuaL Speech rEcognition, *FALSE* is proposed, which manages all the previously mentioned challenges. The main contributions of the thesis are as follows:

1. We demonstrate that a targeted adversarial example exists in the multimodal domain by fooling the AVSR model. To the best of our knowledge, we are the first to simultaneously fool the AVSR model and detection network. The AVSR model is fooled to generate targeted adversarial examples, which may lead to a decrease in the correlation between the two modalities. To prevent the detection network, SyncNet, from detecting the adversarial samples, we fool this network by maintaining the correlation between audio and face videos.

2. We conduct comprehensive experiments using state-of-the-art AVSR model on publicly available Lip Reading in the Wild (LRW) dataset and analyses that attacking either audio or video modality results in perceivable distortions. However, the proposed attack, *FALSE* achieves the desired results by adding small and imperceptible distortions to both modalities.

3. We demonstrate the robustness of our proposed attack by successfully circumventing popular defences while maintaining the imperceptibility of added perturbations with an attack success rate of 100%.

1.4 Organisation of Thesis

The rest of the thesis is organised as follows::

Chapter 2: In this chapter, we provide an outline of the existing works in the field of adversarial attacks and defences for image, audio and audio-visual domain. This section briefly describes well-known AVSR models, adversarial attacks, and the detection networks for AVSR adversarial attacks.

Chapter 3: In this chapter, we presents our proposed attack *FALSE* to fool the AVSR model and detection network. The generation of adversarial examples with the loss function details for the AVSR model and detection network is covered in detail.

Chapter 4: In this chapter, we discuss the experimental results to test the effectiveness of the *FALSE* attack. The introduction of datasets, experimental settings, experimental results and the impact of various input transformation defences on *FALSE* attack.

Chapter 5: In this chapter, we summarise the contributions made in the thesis followed by the future work to be done in this domain.

Chapter 2

Background

2.1 Adversarial Attacks

Deep Learning is making significant progress in solving problems in the field of computer vision [16, 17], speech recognition [18, 19], malware classification [20], natural language processing [21], anomaly detection [22], and many more. However, these models are vulnerable to adversarial attacks based on the imperceptible changes in the input at test time [23, 24]. Adversarial examples are created by adding imperceptible perturbations (or noise) to the input examples with an objective to fool the machine learning models [25]. The adversarial attacks can be classified as targeted or untargeted based on the purpose of the adversary [26]. In untargeted attacks, the aim is to expect any incorrect label on classification. In contrast, targeted attacks create an adversarial example that predicts a particular target label chosen by an adversary. Based on the adversary knowledge, the adversarial attacks can be categorised into the white box, or black box attacks [27]. In a white box setting, the adversary has complete knowledge about the model architecture and parameters, while in a black box setting, the adversary has limited or no knowledge about the model architecture or its parameters.

2.1.1 Adversarial Attacks in the Image Domain

The existing image classification deep learning models recognise the images with near-human accuracy [13]. The authors in [12] proposed the first paper, which shows that image

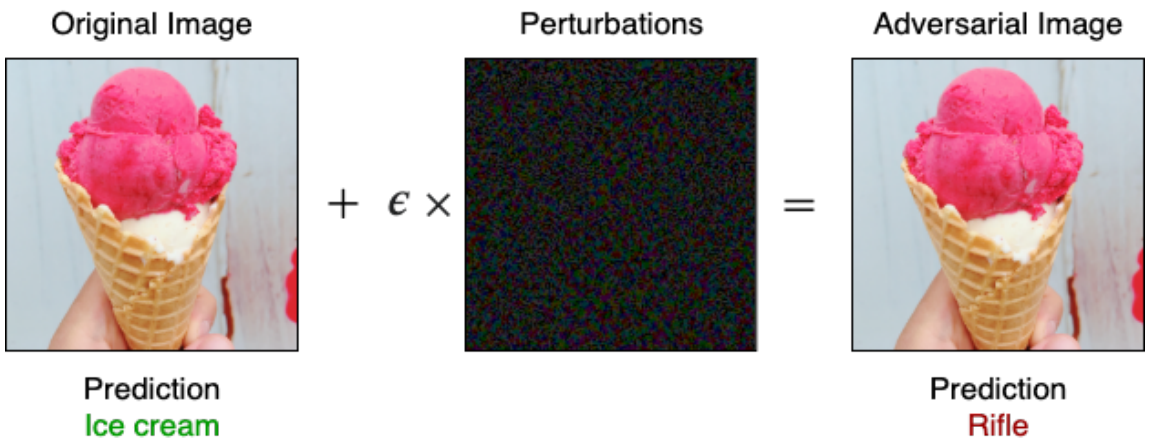

Figure 2.1: Illustration of an adversarial attack on an image classification model.

classification models are vulnerable to adversarial examples. The adversarial image is generated by adding small perturbations to the original image such that the generated image changes the original classification. Fig. 2.1 demonstrates that the original image, when given as input to the image classification model, the prediction is **Ice cream**, but when a small perturbation (noise) is added to the original image, the generated adversarial image prediction changes to **Rifle**. Here, ϵ is used to ensure small perturbation is added to the original image. Several methods are existing in the literature to perform the adversarial attack in the image domain [28]. The methods like Fast Gradient Sign Method (FGSM) [14], Iterative Gradient Sign Method (IGSM) [28], Projected Gradient Descent (PGD) [29], DeepFool [30] and C&W's attack [31] are used to generate adversarial image. The following two commonly used methods FGSM and IGSM, are covered in detail.

1. **Fast Gradient Sign Method (FGSM)**: The method finds the perturbations to be added using the sign of the gradients [14]. The gradients are calculated by computing the derivative of the loss function with respect to the input image. Mathematically,

$$x^{adv} = x - \epsilon * \text{sign}(\nabla_x L(f(x), y)) \tag{2.1}$$

   where, x and x^{adv} are original and adversarial image, y is the target output label, ϵ is a step size to make sure that small perturbation is added, $f(x)$ is the original label, L is loss function and ∇_x represents derivative with respect to x.

2. **Iterative Gradient Sign Method (IGSM)**: The authors in [28] introduces an

improved version of FGSM where the smaller perturbation is added at each iteration. Mathematically, at each iteration n,

$$x_{n+1}^{adv} = x_n^{adv} - \epsilon * \text{sign}(\nabla_x L(f(x_n^{adv}), y))$$
$$\text{such that } x_0^{adv} = x \tag{2.2}$$

where, x_n^{adv} denotes the adversarial image at n^{th} iteration. The adversarial attack can be performed using either black-box or white-box settings as suggested in [14, 12, 24, 32].

2.1.2 Adversarial Attacks in the Audio Domain

Several interesting works are proposed in the field of audio to perform speech to text recognition. The research in performing adversarial attack in the field of the audio domain is limited compared to images. This behaviour is due to the following challenges (i) It is difficult to deal with the information change in the time domain compared to the image domain. (ii) The audio sampling rate is relatively high compared to images having hundred or thousands of pixels [33]. Hence, it is slightly more challenging to generate adversarial audios than images. The adversarial audio is generated by adding carefully crafted perturbations to the original audio such that the generated audio transcribes to any text chosen by the adversary, as demonstrated in Fig. 2.2. Here, ϵ is used to ensure quieter perturbation is introduced while generating adversarial audio.

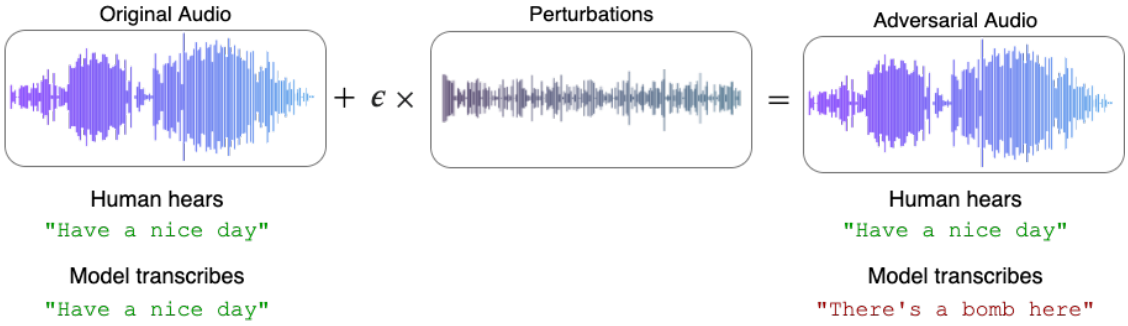

Figure 2.2: Illustration of adversarial attack on the automatic speech recognition model.

The authors in [34] proposed an end-to-end white-box attack iterative and optimisation-based attack, which adds an imperceptible perturbation to the input audio samples. The

authors in [34] use the following optimisation problem :

$$\text{minimise } l(f(a+\delta), y) \quad \text{such that } ||\delta|| < \epsilon \tag{2.3}$$

where, a is original audio, δ is added perturbation to audio sample, $f(.)$ is ASR model, ϵ is used to ensure that the perturbation δ is within a small range, l is loss function and y is the target label. The aim is to minimise the loss function l, which is possible when the ASR model gives the target phrase y as the transcription corresponding to the given audio. The authors in [35] proposed universal adversarial perturbations, which, when added to any audio, will cause the mistranscription by the corresponding speech recognition model. The authors in [36] demonstrate the adversarial attack in devices like Google Home, Amazon's Alexa, Microsoft's Cortana with 98% attack success rate.

2.1.3 Adversarial Attacks in the Audio-Visual Domain

The AVSR models find many applications in the security-critical environment but are vulnerable to adversarial attacks like image and audio modalities. To the best of our knowledge, only one untargeted attack proposed by [15] is available for fooling the AVSR model. Fooling an AVSR model is difficult due to the number of non-trivial challenges like (i) presence of temporal dimension, (ii) presence of non-differentiable layers in the existing AVSR model, (iii) audio and visual modalities complement each other. However, if an adversarial attack is performed on the AVSR model, the detection network can easily detect the generated adversarial example [15] (refer section 2.3).

2.2 AVSR Models

ASR converts utterances to the corresponding transcriptions by taking audio as input [37] and is efficient in establishing an effective interface between human and machine interaction. The efficiency of these model decreases when the audio is distorted by noise [38]. On the other hand, speech recognition can also be done using visual information. The authors designed a visual speech recognition model based on Long Short-Term Memory (LSTM) using visual-only features [39]. Audio-Visual Speech Recognition (AVSR) models seem to be one of the most favourable solutions for speech recognition by utilising both audio and visual modalities. For predicting words, the authors in [40] presented an AVSR model consisting of two streams

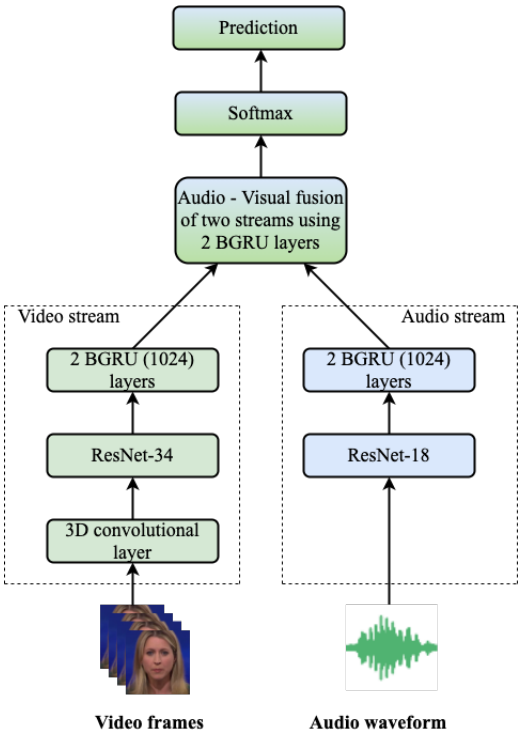

Figure 2.3: Overview of an end-to-end Audio-Visual Speech Recognition (AVSR) model.

that extract features from the mouth region and spectrogram. Each stream consists of an encoder followed by Bidirectional Long Short-Term Memory (BLSTM). The encoder compresses the high dimensional input to low dimension representation. The BLSTM is used to model the temporal dynamics of the features in each stream. Subsequently, the BLSTM outputs of both streams are concatenated and given to another BLSTM to predict the transcription. In addition, the AVSR model proposed in [4] consists of audio streams and visual streams to extract features from audio waveforms and raw images as shown in Fig. 2.2. Each stream consists of ResNet to extract features, followed by a BGRU layer to model temporal dynamics. Subsequently, to fuse the information from both the streams and jointly model temporal dynamics, 2 BGRU layers are added on top of the two streams. Finally, the output of 2 BGRU layers is given to the softmax layer, which assigns a label to each frame. To the best of our knowledge, this AVSR model [1] is the current state-of-the-art

[1]Link to the implementation and pre-trained model: `https://github.com/mpc001/end-to-end-lipreading`

trained on the large publicly available LRW dataset for audio-visual word recognition.

2.3 Detection Network

The authors in [10] proposed a network called as SyncNet [2], which provides the confidence score or correlation between the audio and face videos. This proposed network consists of two streams that take as input the MFCC features of the audio and the extracted mouth region from the face video as shown in Fig. 2.3. Subsequently, the confidence score is computed for a particular offset; the offset is calculated using a sliding window approach. The distance is calculated between one 5-frame video feature and all audio features in the $\pm$ 1 second range for each sample. The offset is found when the distance is minimised. The difference between the minimum and median of the Euclidean distances calculated over all the windows is used to find the confidence score for a certain offset [15].

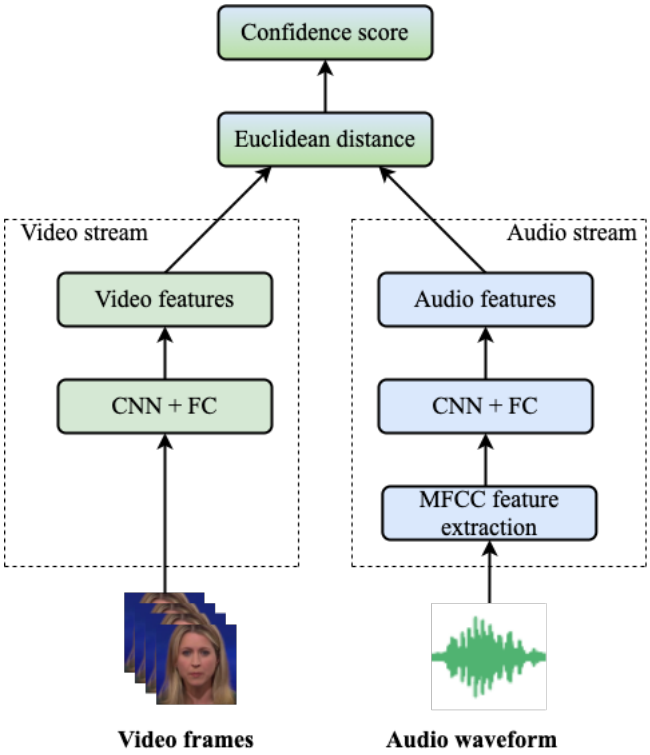

Figure 2.4: Overview of the detection network (SyncNet).

[2]Link to the implementation and pre-trained model: `https://github.com/joonson/syncnet_python`

In the original video sample, the audio and face videos are highly correlated. In contrast, for the adversarial sample, the correlation decreases between the audio and face videos due to the added perturbations [15]. The key idea of the detection method to detect adversarial examples proposed by the authors in [15] is that the correlation between the audio and video streams in an original sample would be higher than the adversarial sample. The detection method uses SyncNet [10] as the detection network to find the correlation. To the best of our knowledge, there exists only one detection method to identify the adversarial examples on the AVSR models [15].

2.4 Adversarial Defences

There has been a significant increase in research in the field of constructing defences to prevent adversarial attacks [27, 41]. While several defences are proposed in the white-box setting but the complete solution has not been found yet [42]. In the image domain, some input transformation defences like bit reduction [43], JPEG-compression [43], box blur [44], median blur [45] etc. are proposed to mitigate the added noise in the clean sample. The input transformations are a widely used method due to their low operation cost and easy integration with the existing architecture [46]. Furthermore, adversarial training a neural network seems to make a more robust machine learning model [47]. However, these defences are more expensive to train [42].

In comparison to the image domain, only a few defences have been proposed in the audio domain [42]. There are some pre-processing defences proposed in the audio domain like local smoothing, downsampling, and quantisation to mitigate the added adversarial perturbations [48]. Furthermore, the authors in [49, 50] use audio-preprocessing methods for the detection of adversarial examples. The authors in [46] proposed a defence method against adversarial attack on the state-of-the-art ASR models, which detects the adversarial example. To detect the adversarial example, the method checks whether the first half of the audio waveform classification is similar to the first half of the complete audio waveform classification [31]. Though, the authors in [31] demonstrated that utilising temporal dependency as suggested in [46] is not effective in detecting the adversarial perturbations in the audio domain [42]. In this thesis, some audio and image-based defences are applied to prove the effectiveness of the proposed attack, which will be discussed in detail in the upcoming chapter 4.

Chapter 3

Proposed Work

3.1 Introduction

This section describes our proposed attack *FALSE* to perform a targeted adversarial attack on the AVSR model while keeping the generated adversarial example undetected. Based on the confidence score, the detection network can identify whether the given video sample is original or adversarial (as discussed in chapter 2.3). We simultaneously attack the AVSR model and detection network to prevent this detection and achieve the required target. The demonstration of the *FALSE* attack is presented in Fig. 3.1 ,which consists of state-of-the-art AVSR model [4] and detection network [10]. The first section discusses how to fool the AVSR model and detection network individually, followed by simultaneously fooling both the AVSR model and detection networks. Finally, the last section covers the implementation details of the *FALSE* attack.

3.2 Fooling AVSR Model

This subsection discusses how to perform an adversarial attack on the AVSR model. The AVSR model $f(V, a)$ takes face video V and audio a as input and predict the word y corresponding to the given input (refer Fig. 3.1). The targeted attack is performed by adding the perturbations to the original inputs a and V. Our aim is to generate adversarial audio $\overline{a}$ that sounds similar to a and adversarial face videos $\overline{V}$ that is visually similar to V and $f(\overline{v}, \overline{a})$ gives target word as prediction, which is different from the original predicted

word. The cross-entropy loss function ℓ_1 is used to fool the AVSR model. Either logits or probabilities can be passed as one of the parameters in the loss function. The unnormalised probability given as input to the softmax function is called logits. The softmax function gives the probabilities of a particular word out of the set of labels as output. The authors in [51] observed that the better way to generate an adversarial example while fooling the deep learning models is to use logits preferably rather than probabilities in the loss function. The experimental results presented in chapter 4 also support this observation. Therefore, we implement the $CrossEntropy$ loss function ℓ_1 using logits, and it can be represented as :

$$\begin{aligned} \ell_1(V, a, y) = CrossEntropy(z, y) &= -\log\left(\frac{\exp(z[\mathrm{y}])}{\sum_j \exp(z[j])}\right) \\ &= -z[\mathrm{y}] + \log\left(\sum_j \exp(z[j])\right) \end{aligned} \tag{3.1}$$

where, V and a are face videos and audio; target label is denoted by y, and the logits obtained corresponding to the given V, a is represented by z.

The proposed targeted attack on the AVSR model is performed using the IGSM method; the IGSM method is discussed in detail in Section 2.1.1. Essentially, the adversarial audio and face videos are created by perturbing the original samples with well-crafted and imperceptible perturbations. The derivatives of the loss function with respect to input audio and face video samples is used to calculate the added perturbations. Specifically, the attack is performed using:

$$V_{n+1} = V_n - \epsilon_V^a * sign(\nabla_V \ell_1(V_n, a_n, y)) \tag{3.2}$$

$$a_{n+1} = a_n - \epsilon_a^a * sign(\nabla_a \ell_1(V_n, a_n, y)) \tag{3.3}$$

where, a_n and V_n denotes the adversarial audio samples and face videos at n^{th} iterations, respectively; the cross-entropy loss is represented using ℓ_1; the logits are obtained from the output of BGRU and is denoted by z; y is target label; ϵ_a^A and ϵ_V^A are step sizes for audio and video modality. Please note that the step sizes are set to small values (for parameter tuning, refer to Section 3.5) such that while generating an adversarial example, the changes made are imperceptible. Furthermore, the value of loss function is minimum, when the the target label y and AVSR model prediction on adversarial examples $f(\overline{V}, \overline{a})$ are same, in other words $f(\overline{V}, \overline{a}) = y$.

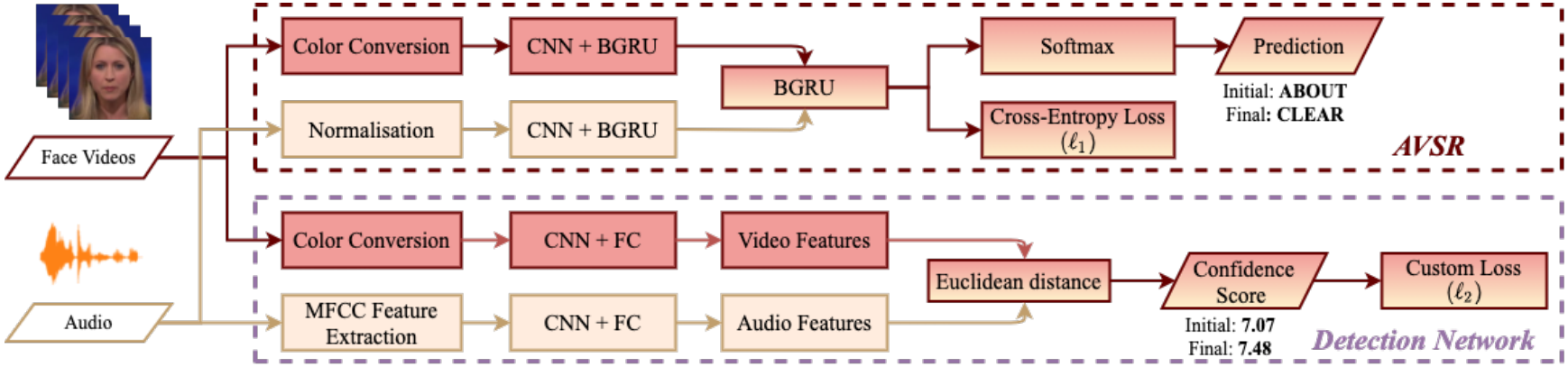

Figure 3.1: Proposed architecture to perform *FALSE* attack.

3.3 Fooling Detection Network

This subsection covers how to perform an adversarial attack on the detection network s. The input audio a and face videos V is given as input to the network s, which calculates the confidence score between the face videos and audio (refer Fig. 3.1). The correlation between the audio and face videos is obtained using the confidence score, which is higher for the original sample than the adversarial sample. The detection network makes use of the confidence score for distinguishing between the original and adversarial examples. To avoid this detection, the detection network is fooled with the goal that the adversarial samples confidence score is always higher than the given original samples confidence score. The custom loss for the detection network is defined for this purpose. The difference between the confidence score of original and adversarial samples is characterised as the custom loss, which is represented by ℓ_2. The loss ℓ_2 is expressed mathematically as:

$$\ell_2(\tau_0, \tau_a) = \max(0, \tau_0 - \tau_a) \tag{3.4}$$

where, τ_a and τ_0 are the confidence scores of the adversarial and original samples, respectively. When the confidence score of adversarial samples is greater than or equal to the original confidence score, the loss ℓ_2 achieves minimum value. The generation of adversarial face videos and audio by keeping the high confidence score is done using the following equations:

$$V_{n+1}^{adv} = V_n - \epsilon_V^s * sign(\nabla_V \ell_2(\tau_n, \tau_o)) \tag{3.5}$$

$$a_{n+1}^{adv} = a_n - \epsilon_a^s * sign(\nabla_a \ell_2(\tau_n, \tau_o)) \tag{3.6}$$

where, the adversarial audio samples and face videos at n^{th} iterations is represented using a_n and V_n respectively; the custom loss function denoted by ℓ_2; the step sizes for audio and face video are ϵ_a^S and ϵ_V^S respectively; τ_n represents the confidence score at n^{th} iteration. The step sizes ϵ_V^S and ϵ_v^A are set to a small value to generate an undetectable adversarial example (For parameter tuning, refer Section 3.5).

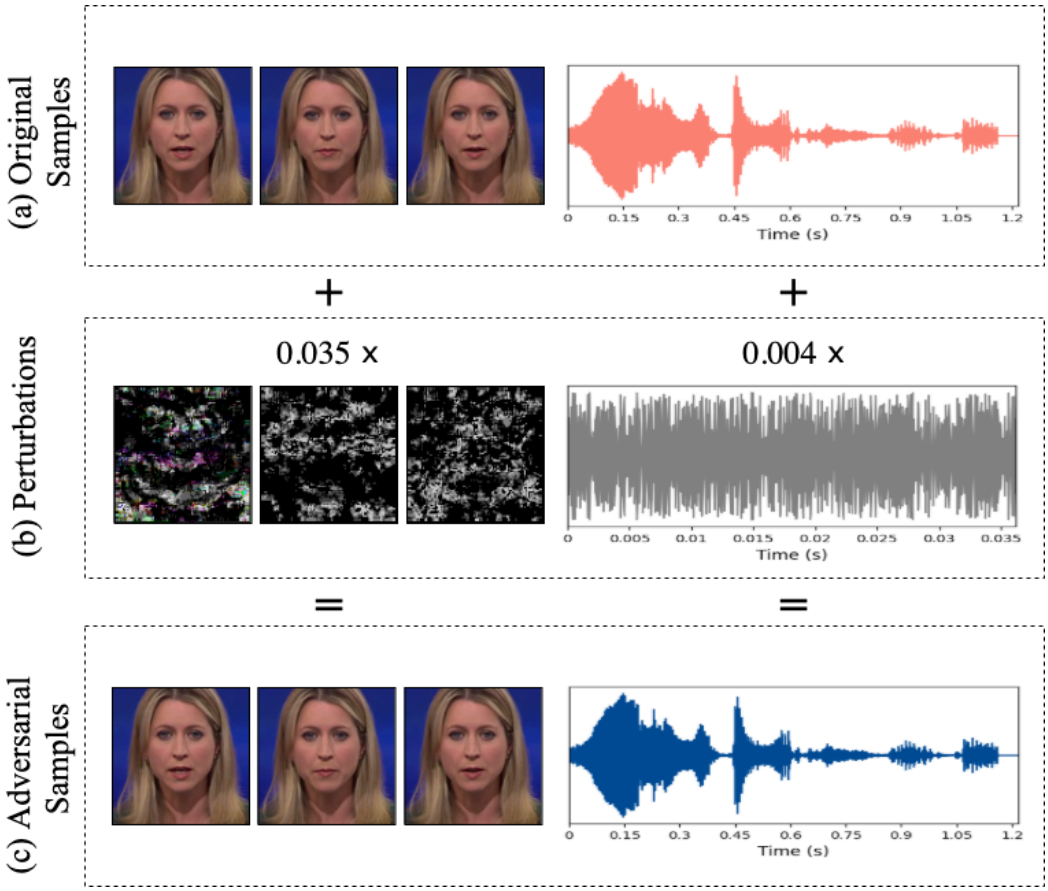

Figure 3.2: Illustration of the generation of adversarial face video and audio using *FALSE* attack.

3.4 FALSE attack: Fooling AVSR and Detection Network

This subsection describes our proposed attack, *FALSE* to generate the adversarial examples. We simultaneously fool the AVSR model and the detection network to achieve the required target while maintaining the correlation between the two modalities. The proposed architecture is shown in Fig. 3.1, which demonstrates the attacking of both the AVSR model and the detection network. To generate adversarial audio and face videos, we alternatively perform the attack on the AVSR model and detection network. Explicitly, we first give original audio and face videos as input to perform the targeted attack on the AVSR model. The adversarial examples are generated using equations (3.2) and (3.3). A detection network

can detect these generated adversarial examples. Therefore, the detection network is fooled for these adversarial examples by increasing their confidence score to prevent such detection. For this, the equations (3.5) and (3.6) are used to perform an adversarial attack on the generated examples with the aim to increase the confidence score. But these generated adversarial examples, when given to the AVSR model, may not predict the required target label. Therefore, we fool both the AVSR model and the detection network simultaneously until the target label is achieved with a high confidence score. The proposed algorithm to perform *FALSE* attack are provided in Algorithm 1.

For visualisation of generated adversarial example, consider Fig. 3.2, which shows the generated adversarial face video and audio example crafted using our *FALSE* attack. The lip-region and complete facial area are given as inputs to the AVSR model and the detection network, respectively. Therefore, the perturbations are added in the entire facial region in the *FALSE* attack. For better visualisation, the scaling is done to show the added imperceptible perturbations in the case of face videos. To visualise the added perturbations in audio, audio perturbations are plotted for a shorter duration (0.035 secs). The original and adversarial samples predict the word as **ABOUT** and **CLEAR**, respectively when given as input to the AVSR model.

3.5 Implementation Details

Both the AVSR model and the detection network take different ranges of input face videos and audio waveform. For example, the pixel intensities range from 0 to 255 for the detection network, while the range is 0 to 1 for the AVSR model. To this end, a scaling layer is added in the preprocessing step to ensure proper inputs to the models. Furthermore, to introduce minimal changes in audio and face videos while generating adversarial audio and face videos, the step sizes (ϵ_a^A and ϵ_v^A) are selected. For the AVSR model, the step size ϵ_v^A is set to 0.00392 (1/255), which is the minimum possible pixel change when the range is 0 to 1. Likewise, ϵ_v^S for the detection network whose input face videos ranges from 0 to 255 is set to 1.0. Although, experiments are performed to find the suitable step size for audio to generate the imperceptible adversarial audio. For the AVSR model, the step size ϵ_a^A is 0.00015 (5/32767), while for the synchronisation-based detection network, ϵ_a^S is 5.0. Moreover, some non-differentiable layers are present during preprocessing, preventing backpropagation till

Algorithm 1 *FALSE* Attack

Require: face videos V; audio a; AVSR model $f(.)$; detection network $s(.)$; step sizes ϵ_V^A and ϵ_a^A for AVSR; and step sizes ϵ_V^S and ϵ_a^S for detection network ; target label y

Ensure: Adversarial face video ($\overline{V}$) and audio ($\overline{a}$)

$\tau_o = s(V, a)$

▷ *Attacking both AVSR and detection network*

while $f(V, a) \neq y$ or $s(V, a) \leqslant \tau_o$ **do**

  while $f(V, a) \neq y$ **do** ▷ *Attacking AVSR*

    $V = V - \epsilon_V^A * \text{sign}\ (\nabla_V \ell_1(V, a, y))$ ▷ *refer eq.* (3.2)

    $a = a - \epsilon_a^A * \text{sign}\ (\nabla_a \ell_1(V, a, y))$ ▷ *refer eq.* (3.3)

  end while

  $\tau_a = s(V, a)$

  while $\tau_a \leqslant \tau_o$ **do** ▷ *Attacking detection network*

    $V = V - \epsilon_V^S * \text{sign}\ (\nabla_V \ell_2(\tau_0, \tau_a))$ ▷ *refer eq.* (3.5)

    $a = a - \epsilon_a^S * \text{sign}\ (\nabla_a \ell_2(\tau_0, \tau_a))$ ▷ *refer eq.* (3.6)

    $\tau_a = s(V, a)$

  end while

end while

$\overline{V} = V$

$\overline{a} = a$

return ($\overline{V}$, $\overline{a}$)

the original audio and face videos. These non-differentiable layers result in gradient masking, which makes it harder for the adversary to perform an adversarial attack [52]. In the case of face videos, this problem is resolved by replacing the non-differentiable layers with their alternative differentiable functions. An open-source computer vision library Kornia provides the differentiable functions [44, 53]. For the AVSR model, the image normalisation and grayscale conversion functions are replaced by their respective differentiable functions. Likewise, the detection network extracts MFCC features of the audio using the feature extraction phase, and this phase consists of some non-differentiable functions. Therefore, the code to extract the MFCC features is written by maintaining the differentiability property

at each layer. It is important to emphasise that the efficacy of the modified model and the original model remains the same.

Chapter 4

Experimental Results

4.1 Dataset

For conducting the experiments, we use the publicly available Lip Reading in the Wild (LRW)[1] dataset [54]. The statistics of the LRW dataset is shown in table 4.1. The test set comprises 25000 video clips, which consists of 29 frames and is 1.16 seconds long. The dataset has been created from broadcast content of BBC News and comprises 1000 utterances of 500 words where the target word is present in the middle of the video. For our experiments, only those samples are used which are correctly classified by the AVSR model. The list of 500 words is shown in Appendix A.

Table 4.1: Statistics of the LRW dataset

Set	**Number of Samples**
Train	488766
Validation	25000
Test	25000

[1]Link to the dataset: `https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html`(For non-commercial individual research and private study use only. BBC content included courtesy of the BBC.)

4.2 Performance Metrics

We use the attack success rate, average audio distortion, and average video distortion as evaluation metrics to evaluate our proposed attacks performance. The metric attack success rate is defined as the rate of an incorrectly classified label when the generated adversarial example is given as input to the AVSR model. Moreover, for video modality, the metric is average video distortion, δ_∞. The video distortion for the individual samples is defined as the maximum change in pixels between adversarial and original face video. Mathematically as suggested in [13],

$$||V - \overline{V}||_\infty = max(|V_1 - \overline{V}_1|, \cdots, |V_n - \overline{V}_n|) \tag{4.1}$$

where, V is the original face videos and $\overline{V}$ is the generated adversarial face videos. Additionally, for audio modality, the metric is average audio distortion, D, which is defined as the relative loudness of the perturbation δ with respect to an original audio a. Mathematically, the distortion $D_{a,\delta}$ is defined as:

$$D_{a,\delta} = dB(\delta) - dB(a) \tag{4.2}$$

where, the decibel value of the audio α is represented by $dB(\alpha)$, as suggested in [34]. The perturbation introduced is quieter than the original audio; therefore, the difference in equation (4.2) is always negative [34]. The smaller the average distortion metric, the quieter the distortion and thus the greater the efficacy.

4.3 Experimental Settings

We perform the attack using $Targeted_1$ and $Targeted_2$ settings to analyse the effectiveness of *FALSE* attack. In the $Targeted_1$ setting, the attack is performed on the AVSR model by setting the target to the second most probable label (the probabilities are given by the AVSR model), and this target label is easier to achieve by adding small adversarial perturbations in the input samples. On the contrary, for the $Targeted_2$ setting, the labels are set to the least probable label given by the AVSR model. In this case, a substantial adversarial perturbation is added to the input sample to generate an adversarial example, making the attack difficult to perform.

4.4 Comparative Analysis

In this section, we will discuss the efficacy of the proposed *FALSE* attack using the evaluation of the experimental results. The proposed *FALSE* attack is compared to audio-only, video-only, and combined loss attacks for a more thorough evaluation. Only one modality is perturbed in video-only and audio-only attacks, with the other modality, remains unchanged. These attacks are carried out to determine the role of each modality in the execution of the attack. Likewise, we devise the combined loss attack to search for an alternate approach to attack the AVSR model and the detection network simultaneously. In the combined loss attack, the adversarial face video and audio are generated using Equations (3.5) and (3.6), with the variation that the loss $\ell_2(\tau_0, \tau_a)$ is substituted with the loss ℓ_3. Where, the loss ℓ_3 is a combination of custom loss ℓ_2 (Section 3.3 for reference) and cross-entropy loss ℓ_1 (Section 3.2 for reference). Mathematically,

$$\ell_3 = c_1 * \ell_1(v, a, y) + c_2 * \ell_2(\tau_o, \tau_a) \tag{4.3}$$

where, c_1 and c_2 are hyperparameters that signify the contribution of each loss functions in the combined loss attack. We conduct several experiments to find the best value of the hyperparameters, which is used to calculate the combined loss. The best values of c_1 and c_2 are found to be 1 and 9.87, respectively. Moreover, the proposed *FALSE* attack is compared with Fooling Audio-VisuaL Speech Recognition using Probabilities (*PFALSE*) and Fooling Audio-VisuaL Speech Recognition by Restricting Video Distortions (*RFALSE*), which are devised from *FALSE* by substituting logits with probability in the loss functions and performing attack by constraining the video distortion of video samples to 5, respectively.

The performance comparison of *FALSE* attack using different approaches is presented in Table 4.2. Kindly note that all these attacks are implemented by keeping the correlation between the two modalities, such that the generated adversarial examples remain unrecognised by the detection network. It can be inferred from the table that performing attacks on both audio and video modalities while generating adversarial examples is better than performing an attack on a single modality. The reason for this is that large perturbations are required if adversarial perturbations are added in only one modality (in particular, video-only and audio-only attack). Furthermore, It can be inferred from the table that the performance of the *FALSE* attack is better than the combined loss attack, as *FALSE* attack fools the AVSR model by adding fewer distortions in both modalities. Additionally, it can be observed that

Table 4.2: Comparison of the proposed *FALSE* attack with different approaches

	Targeted$_1^*$			**Targeted$_2^*$**		
Attack Types	**Attack Success Rate (in %)**	**Average Video$^+$ Distortion, δ_∞**	**Average Audio$^+$ Distortion, D (in dB)**	**Attack Success Rate (in %)**	**Average Video$^+$ Distortion, δ_∞**	**Average Audio$^+$ Distortion, D (in dB)**
Audio-only	100.0	–	-46.89	100.0	–	-27.87
Video-only	100.0	3.84	––	100.0	46.69	––
Combined Loss	100.0	2.97	-56.89	100.0	10.52	-45.53
RFALSE[1]	99.19	2.68	-55.27	20.45	4.82	-38.93
PFALSE[2]	100.0	2.83	-52.27	35.23	18.16	-33.71
FALSE	**100.0**	**2.74**	**-57.42**	**100.0**	**8.73**	**-46.18**

*: The target label is set to the second-most probable label in $Targeted_1$ setting and least probable label in $Targeted_2$ setting, respectively.

+: The smaller the value of δ_∞ or D, the better the imperceptibility and, as a result, the better the performance.

−: Audio-only attacks are carried out by adding the perturbations in the audio while leaving the face videos unaffected.

−−: Video-only attacks are carried out by adding the perturbations in the face videos while leaving the audio unaffected.

1: *RFALSE* is devised by constraining δ_∞ to 5 in *FALSE* attack.

2: *PFALSE* is devised by substituting logits with probability in the loss functions and by constraining δ_∞ to 20 and D to -30 dB respectively in the *FALSE* attack.

Note : The audio-visual samples that the AVSR model classifies correctly are used to perform the attacks.

FALSE significantly outperforms *PFALSE*, which indicates that there is performance degradation when probabilities as a parameter in the loss function replace logits. This observation is in accordance with the conclusion made in [51]. The *PFALSE* attack restricts the video distortion δ_∞ to 20 pixels and audio distortion D to -30 dB while performing the attack. As shown in Table 4.2 for $Targeted_2$ setting, this attack introduces higher distortions in audio and face videos, due to which the attack success rate is very less.

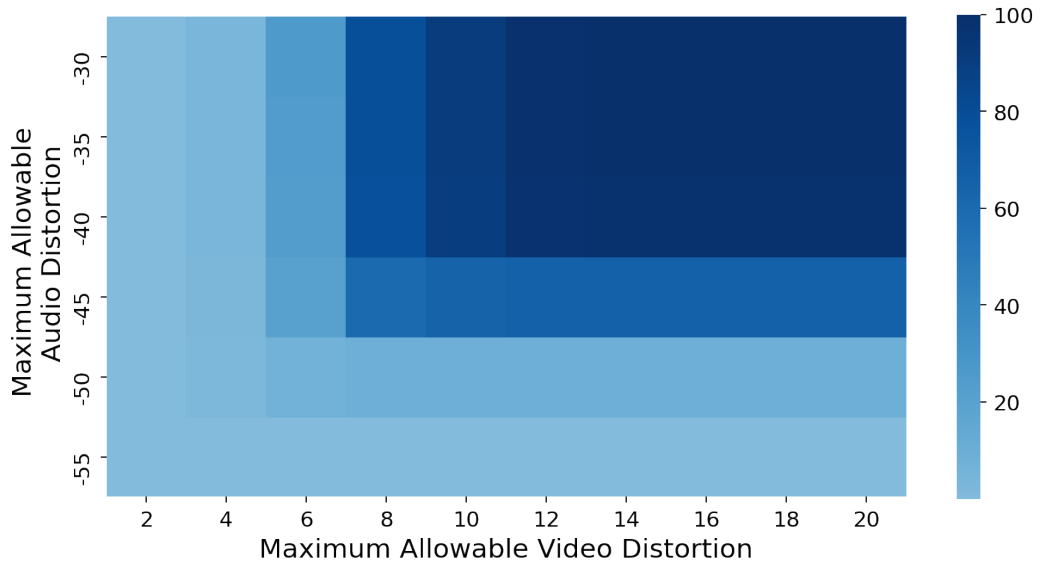

Figure 4.1: Heatmap representation of *FALSE* attack success rate in $Targeted_2$ setting with x-axis representing maximum allowable distortion in video (δ_∞) and y-axis represents audio (D) distortion.

Our proposed *FALSE* attack introduces less average audio and video distortions by achieving a 100% targeted attack success rate. Normally, the video distortion is set to 5 pixels. However, it is not possible to fool the AVSR model for all the target labels by restricting video distortion to 5 pixels. Therefore, in *RFALSE* attack, the attack success rate is less than 100%. This outcome is more noticeable when the attack is performed in the $Targeted_2$ setting, as it requires large perturbations and is challenging to perform. Reducing step-sizes like setting the step-size less than 1 in video modality makes it possible to get a 100% attack success rate. Although, in our thesis, we avoid this reduction of step sizes because due to quantisation error, it will generate adversarial examples which cannot be saved and reused back to fool the AVSR models later. The visualisation of the attack success rate *FALSE* by restricting the maximum values of δ_∞ and D is presented in Fig. 4.1 using a heatmap. The heatmap represents the efficacy of the proposed *FALSE* attack in $Targeted_2$ setting.

4.5 Evaluation of defences on FALSE attack

As described in section 2.4, it is shown that adversarial defences can easily mitigate several powerful attacks due to which the adversarial attacks are of limited applicability [55]. For a more thorough understanding, we evaluate our proposed *FALSE* attack against the popular input transformation defences on audio and videos modalities. In comparison to the image and audio domain, adversarial attacks and defences are not widely studied for the video modality. Hence, we evaluate the attack using the following four popular image-based defences that can be used for the video modality:

1. **Bit Reduction:** It removes small (adversarial) variations in pixel values from an image by performing a simple type of quantisation [56]. As suggested in [43], from 8 to 5 bits, the bit frames are reduced in our experiments.

2. **JPEG-Compression:** For our experiments, compression at quality level 75 (out of 100) is done in the face videos [43].

3. **Box Blur:** It uses a box filter to blur an image; by replacing each pixel of an image with the average of its neighbouring pixels, [44]. In our experiments, the blurring kernel size is taken as 3×3 [44].

4. **Median Blur:** It replaces the central pixel with the median of the neighbouring pixels. [45]. In our experiments, the blurring kernel size is taken as 3×3 [44].

Additionally, we evaluate the attack using the following three popular audio-based defences:

1. **MP3 Compression:** It is preprocessing defence methods that mitigate the effect of audio adversarial examples. In our experiments, the input audio is compressed at a constant bit rate of 48kbps [57, 58].

2. **Re-sampling:** The sampling rate of the original audio used in our experiments is 16kHz, the input audio is re-sampled to 8kHz, and again re-sample the audio back to the actual sampling rate, i.e. 16kHz [57].

3. **White noise addition:** It is a conventional digital distortion, which is added at a signal to noise ratio (SNR) of 60 dB to the audio [57].

Table 4.3: Impact of combination of audio and image defence on the proposed *FALSE* attack.

		Targeted$_1^*$		**Targeted$_2^*$**	
Defence Used	**Model$^{\#}$ Accuracy (in %)**	**Average Video$^+$ Distortion, δ_∞**	**Average Audio$^+$ Distortion, D (in dB)**	**Average Video$^+$ Distortion, δ_∞**	**Average Audio$^+$ Distortion, D (in dB)**
None (Proposed Attack)	**98.38**	**2.74**	**-57.42**	**8.73**	**-46.18**
Bit Reduction (BR)	96.60	3.69	-56.43	11.99	-44.98
JPEG-Compression (JC)	96.60	3.85	-56.13	12.32	-44.73
Box Blur (BB)	96.60	2.97	-56.89	10.52	-45.53
Median Blur (MB)	96.40	3.01	-56.70	11.08	-45.01
MP3-Compression	93.20	2.95	-11.94	17.67	-10.95
Re-sampling	90.40	2.85	-56.85	11.18	-45.16
White Noise	96.60	2.89	-50.00	10.02	-43.08
BR + MP3-Compression	93.00	4.67	-11.25	10.03	-9.65
JC + MP3-Compression	93.20	5.04	-11.06	25.61	-9.39
BB + MP3-Compression	93.20	3.32	-11.62	19.24	-10.48
MB + MP3-Compression	93.40	3.51	-11.81	20.76	-10.51
BR + Re-sampling	90.40	3.39	-57.80	14.06	-43.98
JC + Re-sampling	90.60	3.51	-57.39	14.56	-43.66
BB + Re-sampling	90.20	2.95	-56.61	11.75	-44.72
MB + Re-sampling	90.20	2.94	-56.39	12.60	-44.15
BR + White Noise	96.60	3.72	-50.18	11.91	-42.32
JC + White Noise	96.60	3.85	-50.01	12.31	-42.05
BB + White Noise	96.60	2.98	-50.49	10.49	-42.75
MB + White Noise	96.40	3.03	-50.48	11.09	-42.37

*: The target label is set to the second-most probable label in $Targeted_1$ setting and least probable label in $Targeted_2$ setting, respectively.

$^+$: The smaller the value of δ_∞ or D, the better the imperceptibility and, as a result, the better the performance.

$^{\#}$: The percentage of original audio-visual samples that the AVSR model classifies correctly.

Note: The audio-visual samples that the AVSR model classifies correctly are used to perform the attacks.

Table 4.3 represents the performance of our *FALSE* attack when different feasible combinations of audio and video defences are applied. As suggested by [52] the Backward Pass Differentiable Approximation (BPDA) is used to approximate derivatives of some non-differentiable defences. It can be inferred from the table that the *FALSE* attack can successfully circumvent (bypass) the popular input transformation defences. In addition, when the defences are applied, the AVSR model accuracy decreases slightly. Also, when the input transformation defences are utilised, there is an increase in average video and audio distortions because the added defences disrupt the adversarial perturbations. Furthermore, as can be inferred from Table 4.3, the proposed *FALSE* attack can easily handle the Box Blur and Re-sampling audio defences for video and audio modality, respectively. Although applying JPEG-Compression and MP3-Compression in video and audio is the most challenging defence for the AVSR model because these defences introduce the significant value of distortion in the generated adversarial samples.

4.6 Discussion

The adversarial attack on the AVSR model proposed in [15] uses IGSM to generate the adversarial examples. In the $Targeted_1$ setting, the video and audio distortion of 1.99 and -30.54dB are achieved, while in the $Targeted_2$ setting, it achieves the video and audio distortion of 13.43 and -15.38dB, respectively. The Attacking only Audio-Visual Speech Recognition Model (AAVSR) attack is performed by attacking only the AVSR model, avoiding the attack on the detection network in *FALSE*, to compare the results of the attack proposed in the [15]. For the AAVSR attack, the average video and audio distortions are 1.87 and -60.64 dB in the $Targeted_1$ setting. The average video and audio distortions for the $Targeted_2$ settings are 8.86 and -47.16 dB, respectively. As observed from the results, the AAVSR attack performs better than the attack proposed in [15]. The reason for this is the AAVSR loss function makes use of logits instead of probabilities. Furthermore, in AAVSR, the step size for audio is small, resulting in small perturbations to be added at each iteration. Although, a detection network can easily detect both the attacks proposed in [15] and AAVSR. However, the distortion introduced in the *FALSE* attack is more significant than AAVSR; we still prefer our *FALSE* attack because the detection network cannot detect the generated audio-visual adversarial examples using the proposed algorithm.

It has been observed that it is not possible to perform an untargeted attack using Equations (3.5) and (3.6) by minimising only $\ell_2(\tau_0, \tau_a)$ loss because at each iteration the confidence score increases which signifies the increase in the correlation between the two modalities. As the two modalities are more correlated the prediction at each iteration will remain the same as the initial prediction. The generated adversarial audio and face videos using the Equations (3.5) and (3.6) are not able to misguide the detection network.

The *FALSE* attack is a generic attack; with a slight modification in the proposed algorithm, the proposed attack can fool any alternative AVSR models and detection networks. By utilising the gradient-based attack methods, it is easier to fool the AVSR model. As stated in Section 3.5 for the gradient-based method, If non-differentiable layers prevent gradient backpropagation, replace those with differentiable layers to perform the attack. Likewise, It is easier to get the confidence score as output from any of the detection networks. The proposed custom-based loss function will be exploited directly to fool these detection networks while keeping the two modalities correlated. To the best of our knowledge, for the publically available LRW dataset, there exists no other pre-trained model. Therefore, all the experiments are performed on the AVSR model (refer section 2.2), which is currently state-of-art for the LRW dataset.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Research in the Audio-Visual Speech Recognition (AVSR) model has seen significant progress in recent years. AVSR models find numerous application in different domains. The vulnerability of AVSR models to adversarial examples leads to several uncertain issues and security problems, which motivates us to devise a *FALSE* attack by studying adversarial attacks and defences on these models. As audio and video modalities complement each other, it is more challenging to fool the AVSR model. Furthermore, there is a detection network to identify whether the given video is original or an adversarial one. In this thesis, we are the first one to propose an end-to-end adversarial attack on the AVSR model by maintaining the correlation between the audio and face video samples. The added perturbations are imperceptible while generating an adversarial example. Extensive experiments to fool the AVSR model demonstrate that attacking using both the modalities, i.e. adding perturbation in both modalities, leads to fewer distortions in audio and face videos than attacking either only audio or video modality. Our experiments demonstrate that we can easily surpass the popular input transformation audio and image defences. While creating an AVSR model, the motive is to improve the model accuracy rather than the robustness, due to which there is a possibility to generate adversarial examples that can easily fool the AVSR model. Our findings in this thesis can help understand the requirement to design a secure, reliable and robust AVSR model without reducing the efficiency.

5.2 Future Work

There are various new research directions to work in the field of adversarial attacks on the Audio-Visual Speech Recognition (AVSR) model. Following are the major direction in which we can extend this work :

1. **Study of Defences to make the existing AVSR model more robust** : One of the severe issues with using AVSR in safety-critical contexts is the vulnerability of the AVSR model to adversarial examples. A defence method should be proposed to develop a secure, robust, and trustworthy AVSR model. There exist several defences in the literature in the audio and image domain but it has been observed that sometimes by using the defence, the model accuracy on clean samples decreases and these defences can be circumvented using adaptive attacks [31]. The following defences can be added on the AVSR model to make it more robust:

   (a) Adversarial training in which a network is trained on adversarial examples is one of the few defences against adversarial attacks that withstands strong attacks. The adversarial training defence can be used to make the AVSR model robust against the proposed attack.

   (b) The *FALSE* attack is proposed on the CNN-based models which exploit texture to make a decision rather than on shape. Recent research in the field of transformers demonstrates that the transformers mainly focus on shapes rather than texture to outperform CNN-based models and achieve human-level performance [59]. Hence, we believe that transformers can provide an inherent defence.

2. **Black-Box Adversarial attack on AVSR model**: A white-box or black-box attack can be used to generate adversarial examples. In the black-box approach, the adversary relies on the query to retrieve the information required to generate the adversarial example. To this end, gradient estimation using the finite difference method is commonly used in the image and audio domain to estimate the gradients. Then use iterative gradient sign method (IGSM) and Projected Gradient Descent (PGD) method to generate adversarial examples. There are several interesting works on image, audio and text that have been done using the black-box attack, and that work can be extended for the audio-visual domain.

Bibliography

[1] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," *Issues in Visual and Audio-Visual Speech Processing*, vol. 22, p. 23, 2004.

[3] W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *The journal of the acoustical society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[4] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[5] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: a Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[6] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust Audio-Visual Speech Recognition under Noisy Audio-Video Conditions," *IEEE transactions on cybernetics*, vol. 44, no. 2, pp. 175–184, 2013.

[7] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual Biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006.

[8] P. Borde, A. Varpe, R. Manza, and P. Yannawar, "Recognition of Isolated Words using Zernike and MFCC Features for Audio Visual Speech Recognition," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 167–175, 2015.

[9] M. Cristani, M. Bicego, and V. Murino, "Audio-Visual Event Recognition in Surveillance Video Sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[10] J. S. Chung and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," in *Workshop on Multi-view Lip-reading, Asian Conference on Computer Vision (ACCV)*, 2016, pp. 251–263.

[11] J. S. Chung, A. Jamaludin, and A. Zisserman, "You Said That?" in *British Machine Vision Conference (BMVC)*. BMVA Press, 2017.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[13] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2017, pp. 39–57.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] P. Ma, S. Petridis, and M. Pantic, "Detecting Adversarial Attacks on Audio-Visual Speech Recognition," *arXiv preprint arXiv:1912.08639*, 2019.

[16] R. K. Sinha, R. Pandey, and R. Pattnaik, "Deep learning for computer vision tasks: a review," *arXiv preprint arXiv:1804.03928*, 2018.

[17] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.

[18] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer, 2019, vol. 84.

[19] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition using Deep Neural Networks: A Systematic Review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[20] M. Kalash, M. Rochan, N. Mohammed, N. D. Bruce, Y. Wang, and F. Iqbal, "Malware Classification with Deep Convolutional Neural Networks," in *2018 9th IFIP international conference on new technologies, mobility and security (NTMS)*. IEEE, 2018, pp. 1–5.

[21] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[22] R. Chalapathy and S. Chawla, "Deep Learning for Anomaly Detection: A Survey," *arXiv preprint arXiv:1901.03407*, 2019.

[23] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion Attacks Against Machine Learning at Test Time," in *Joint European conference on machine learning and knowledge discovery in databases (ECML PKDD)*. Springer, 2013, pp. 387–402.

[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," *arXiv preprint arXiv:1605.07277*, 2016.

[25] P. Gupta and E. Rahtu, "Mlattack: Fooling semantic segmentation networks by multi-layer attacks," in *German Conference on Pattern Recognition (GCPR)*. Springer, 2019, pp. 401–413.

[26] P. Rathore, A. Basak, S. H. Nistala, and V. Runkana, "Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[27] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial Attacks and Defences: A Survey," *arXiv preprint arXiv:1810.00069*, 2018.

[28] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial Examples in the Physical World," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017.

[29] A. Ilyas, L. Engstrom, and A. Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," *arXiv preprint arXiv:1807.07978*, 2018.

[30] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[31] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On Adaptive Attacks to Adversarial Example defenses," *arXiv preprint arXiv:2002.08347*, 2020.

[32] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks Against Machine Learning," in *ACM Asia conference on computer and communications security (CCS)*, 2017, pp. 506–519.

[33] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems," in *15th ACM Asia Conference on Computer and Communications Security (CCS), Taipei, Taiwan, October 5-9, 2020*. ACM, 2020, pp. 357–369.

[34] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in *IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[35] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. McAuley, and F. Koushanfar, "Universal Adversarial Perturbations for Speech Recognition Systems," *arXiv preprint arXiv:1905.03828*, 2019.

[36] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang, "Devil's whisper: A general Approach for Physical Adversarial Attacks against Commercial Black-Box Speech Recognition Devices," in *29th USENIX Security Symposium*. USENIX Association, 2020, pp. 2667–2684.

[37] A. Graves and N. Jaitly, "Towards End-to-End Speech Recognition with Recurrent Neural Networks," in *International conference on machine learning (ICML)*. PMLR, 2014, pp. 1764–1772.

[38] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[39] S. Petridis, Z. Li, and M. Pantic, "End-to-End Visual Speech Recognition with LSTMs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2592–2596.

[40] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End Audiovisual Fusion with LSTMs," in *Auditory-Visual Speech Processing (AVSP)*. ISCA, 2017, pp. 36–40.

[41] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[42] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and Mitigating Audio Adversarial Examples," *arXiv preprint arXiv:2103.03344*, 2021.

[43] C. Guo, M. Rana, M. Cissé, and L. van der Maaten, "Countering Adversarial Images using Input Transformations," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[44] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski, "Kornia: An Open Source Differentiable Computer Vision Library for Pytorch," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2020, pp. 3674–3683.

[45] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *Annual Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2018.

[46] Z. Yang, B. Li, P.-Y. Chen, and D. Song, "Characterizing Audio Adversarial Examples using Temporal Dependency," *arXiv preprint arXiv:1809.10875*, 2018.

[47] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[48] Z. Yang, B. Li, P.-Y. Chen, and D. Song, "Towards Mitigating Audio Adversarial Perturbations," *Openreview.net*, 2018.

[49] K. Rajaratnam, K. Shah, and J. Kalita, "Isolated and Ensemble Audio Preprocessing Methods for Detecting Adversarial Examples Against Automatic Speech Recognition," *arXiv preprint arXiv:1809.04397*, 2018.

[50] H. Kwon, H. Yoon, and K.-W. Park, "POSTER: Detecting Audio Adversarial Example through Audio Modification," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2019, pp. 2521–2523.

[51] H. Chen, H. Zhang, P. Chen, J. Yi, and C. Hsieh, "Attacking Visual Language Grounding with Adversarial Examples: A Case Study on Neural Image Captioning," in *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018, pp. 2587–2597.

[52] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 274–283.

[53] E. Riba, M. Fathollahi, W. Chaney, E. Rublee, and G. Bradski, "Torchgeometry: When PyTorch meets Geometry," 2018.

[54] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2016, pp. 87–103.

[55] P. Gupta and E. Rahtu, "CIIDefence: Defeating Adversarial Attacks by Fusing Class-Specific Image Inpainting and Image Denoising," in *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6708–6717.

[56] Y. Zeng, H. Qiu, G. Memmi, and M. Qiu, "A Data Augmentation-based Defense Method Against Adversarial Attacks in Neural Networks," in *International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*. Springer, 2020, pp. 274–289.

[57] V. Subramanian, E. Benetos, and M. B. Sandler, "Robustness of Adversarial Attacks in Sound Event Classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, p. 239–243.

[58] N. Das, M. Shanbhogue, S.-T. Chen, L. Chen, M. E. Kounavis, and D. H. Chau, "Adagio: Interactive experimentation with Adversarial Attack and Defence for Audio," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer, 2018, pp. 677–681.

[59] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *arXiv preprint arXiv:2105.10497*, 2021.

Appendix A

List of Classes

1. ABOUT
2. ABSOLUTELY
3. ABUSE
4. ACCESS
5. ACCORDING
6. ACCUSED
7. ACROSS
8. ACTION
9. ACTUALLY
10. AFFAIRS
11. AFFECTED
12. AFRICA
13. AFTER
14. AFTERNOON
15. AGAIN
16. AGAINST
17. AGREE
18. AGREEMENT
19. AHEAD
20. ALLEGATIONS
21. ALLOW
22. ALLOWED
23. ALMOST
24. ALREADY
25. ALWAYS
26. AMERICA
27. AMERICAN
28. AMONG
29. AMOUNT
30. ANNOUNCED
31. ANOTHER
32. ANSWER
33. ANYTHING
34. AREAS
35. AROUND
36. ARRESTED
37. ASKED
38. ASKING
39. ATTACK
40. ATTACKS
41. AUTHORITIES
42. BANKS
43. BECAUSE
44. BECOME
45. BEFORE
46. BEHIND
47. BEING
48. BELIEVE
49. BENEFIT
50. BENEFITS
51. BETTER
52. BETWEEN
53. BIGGEST
54. BILLION
55. BLACK
56. BORDER
57. BRING
58. BRITAIN
59. BRITISH
60. BROUGHT

61. BUDGET
62. BUILD
63. BUILDING
64. BUSINESS
65. BUSINESSES
66. CALLED
67. CAMERON
68. CAMPAIGN
69. CANCER
70. CANNOT
71. CAPITAL
72. CASES
73. CENTRAL
74. CERTAINLY
75. CHALLENGE
76. CHANCE
77. CHANGE
78. CHANGES
79. CHARGE
80. CHARGES
81. CHIEF
82. CHILD
83. CHILDREN
84. CHINA
85. CLAIMS
86. CLEAR
87. CLOSE
88. CLOUD
89. COMES
90. COMING
91. COMMUNITY
92. COMPANIES
93. COMPANY
94. CONCERNS
95. CONFERENCE
96. CONFLICT
97. CONSERVATIVE
98. CONTINUE
99. CONTROL
100. COULD
101. COUNCIL
102. COUNTRIES
103. COUNTRY
104. COUPLE
105. COURSE
106. COURT
107. CRIME
108. CRISIS
109. CURRENT
110. CUSTOMERS
111. DAVID
112. DEATH
113. DEBATE
114. DECIDED
115. DECISION
116. DEFICIT
117. DEGREES
118. DESCRIBED
119. DESPITE
120. DETAILS
121. DIFFERENCE
122. DIFFERENT
123. DIFFICULT
124. DOING
125. DURING
126. EARLY
127. EASTERN
128. ECONOMIC
129. ECONOMY
130. EDITOR
131. EDUCATION
132. ELECTION
133. EMERGENCY
134. ENERGY
135. ENGLAND
136. ENOUGH
137. EUROPE
138. EUROPEAN
139. EVENING
140. EVENTS
141. EVERY
142. EVERYBODY
143. EVERYONE
144. EVERYTHING
145. EVIDENCE
146. EXACTLY
147. EXAMPLE
148. EXPECT
149. EXPECTED
150. EXTRA
151. FACING
152. FAMILIES
153. FAMILY

154. FIGHT
155. FIGHTING
156. FIGURES
157. FINAL
158. FINANCIAL
159. FIRST
160. FOCUS
161. FOLLOWING
162. FOOTBALL
163. FORCE
164. FORCES
165. FOREIGN
166. FORMER
167. FORWARD
168. FOUND
169. FRANCE
170. FRENCH
171. FRIDAY
172. FRONT
173. FURTHER
174. FUTURE
175. GAMES
176. GENERAL
177. GEORGE
178. GERMANY
179. GETTING
180. GIVEN
181. GIVING
182. GLOBAL
183. GOING
184. GOVERNMENT
185. GREAT
186. GREECE
187. GROUND
188. GROUP
189. GROWING
190. GROWTH
191. GUILTY
192. HAPPEN
193. HAPPENED
194. HAPPENING
195. HAVING
196. HEALTH
197. HEARD
198. HEART
199. HEAVY
200. HIGHER
201. HISTORY
202. HOMES
203. HOSPITAL
204. HOURS
205. HOUSE
206. HOUSING
207. HUMAN
208. HUNDREDS
209. IMMIGRATION
210. IMPACT
211. IMPORTANT
212. INCREASE
213. INDEPENDENT
214. INDUSTRY
215. INFLATION
216. INFORMATION
217. INQUIRY
218. INSIDE
219. INTEREST
220. INVESTMENT
221. INVOLVED
222. IRELAND
223. ISLAMIC
224. ISSUE
225. ISSUES
226. ITSELF
227. JAMES
228. JUDGE
229. JUSTICE
230. KILLED
231. KNOWN
232. LABOUR
233. LARGE
234. LATER
235. LATEST
236. LEADER
237. LEADERS
238. LEADERSHIP
239. LEAST
240. LEAVE
241. LEGAL
242. LEVEL
243. LEVELS
244. LIKELY
245. LITTLE
246. LIVES

247. LIVING
248. LOCAL
249. LONDON
250. LONGER
251. LOOKING
252. MAJOR
253. MAJORITY
254. MAKES
255. MAKING
256. MANCHESTER
257. MARKET
258. MASSIVE
259. MATTER
260. MAYBE
261. MEANS
262. MEASURES
263. MEDIA
264. MEDICAL
265. MEETING
266. MEMBER
267. MEMBERS
268. MESSAGE
269. MIDDLE
270. MIGHT
271. MIGRANTS
272. MILITARY
273. MILLION
274. MILLIONS
275. MINISTER
276. MINISTERS
277. MINUTES
278. MISSING
279. MOMENT
280. MONEY
281. MONTH
282. MONTHS
283. MORNING
284. MOVING
285. MURDER
286. NATIONAL
287. NEEDS
288. NEVER
289. NIGHT
290. NORTH
291. NORTHERN
292. NOTHING
293. NUMBER
294. NUMBERS
295. OBAMA
296. OFFICE
297. OFFICERS
298. OFFICIALS
299. OFTEN
300. OPERATION
301. OPPOSITION
302. ORDER
303. OTHER
304. OTHERS
305. OUTSIDE
306. PARENTS
307. PARLIAMENT
308. PARTIES
309. PARTS
310. PARTY
311. PATIENTS
312. PAYING
313. PEOPLE
314. PERHAPS
315. PERIOD
316. PERSON
317. PERSONAL
318. PHONE
319. PLACE
320. PLACES
321. PLANS
322. POINT
323. POLICE
324. POLICY
325. POLITICAL
326. POLITICIANS
327. POLITICS
328. POSITION
329. POSSIBLE
330. POTENTIAL
331. POWER
332. POWERS
333. PRESIDENT
334. PRESS
335. PRESSURE
336. PRETTY
337. PRICE
338. PRICES
339. PRIME

340. PRISON
341. PRIVATE
342. PROBABLY
343. PROBLEM
344. PROBLEMS
345. PROCESS
346. PROTECT
347. PROVIDE
348. PUBLIC
349. QUESTION
350. QUESTIONS
351. QUITE
352. RATES
353. RATHER
354. REALLY
355. REASON
356. RECENT
357. RECORD
358. REFERENDUM
359. REMEMBER
360. REPORT
361. REPORTS
362. RESPONSE
363. RESULT
364. RETURN
365. RIGHT
366. RIGHTS
367. RULES
368. RUNNING
369. RUSSIA
370. RUSSIAN
371. SAYING
372. SCHOOL
373. SCHOOLS
374. SCOTLAND
375. SCOTTISH
376. SECOND
377. SECRETARY
378. SECTOR
379. SECURITY
380. SEEMS
381. SENIOR
382. SENSE
383. SERIES
384. SERIOUS
385. SERVICE
386. SERVICES
387. SEVEN
388. SEVERAL
389. SHORT
390. SHOULD
391. SIDES
392. SIGNIFICANT
393. SIMPLY
394. SINCE
395. SINGLE
396. SITUATION
397. SMALL
398. SOCIAL
399. SOCIETY
400. SOMEONE
401. SOMETHING
402. SOUTH
403. SOUTHERN
404. SPEAKING
405. SPECIAL
406. SPEECH
407. SPEND
408. SPENDING
409. SPENT
410. STAFF
411. STAGE
412. STAND
413. START
414. STARTED
415. STATE
416. STATEMENT
417. STATES
418. STILL
419. STORY
420. STREET
421. STRONG
422. SUNDAY
423. SUNSHINE
424. SUPPORT
425. SYRIA
426. SYRIAN
427. SYSTEM
428. TAKEN
429. TAKING
430. TALKING
431. TALKS
432. TEMPERATURES

433. TERMS
434. THEIR
435. THEMSELVES
436. THERE
437. THESE
438. THING
439. THINGS
440. THINK
441. THIRD
442. THOSE
443. THOUGHT
444. THOUSANDS
445. THREAT
446. THREE
447. THROUGH
448. TIMES
449. TODAY
450. TOGETHER
451. TOMORROW
452. TONIGHT
453. TOWARDS
454. TRADE
455. TRIAL
456. TRUST
457. TRYING
458. UNDER
459. UNDERSTAND
460. UNION
461. UNITED
462. UNTIL
463. USING
464. VICTIMS
465. VIOLENCE
466. VOTERS
467. WAITING
468. WALES
469. WANTED
470. WANTS
471. WARNING
472. WATCHING
473. WATER
474. WEAPONS
475. WEATHER
476. WEEKEND
477. WEEKS
478. WELCOME
479. WELFARE
480. WESTERN
481. WESTMINSTER
482. WHERE
483. WHETHER
484. WHICH
485. WHILE
486. WHOLE
487. WINDS
488. WITHIN
489. WITHOUT
490. WOMEN
491. WORDS
492. WORKERS
493. WORKING
494. WORLD
495. WORST
496. WOULD
497. WRONG
498. YEARS
499. YESTERDAY
500. YOUNG

Publications

1. **Saumya Mishra**, Anup Kumar Gupta, Puneet Gupta, "DARE: Deceiving Audio-Visual Speech Recognition Model", Knowledge-Based Systems, Elsevier (Accepted)