Knowledge-Based Systems 232 (2021) 107503

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

DARE: Deceiving Audio–Visual speech Recognition model

Saumya Mishra[1], Anup Kumar Gupta*[,1], Puneet Gupta

Department of Computer Science and Engineering, IIT Indore, Indore, India

ARTICLE INFO

Article history:
Received 13 April 2021
Received in revised form 19 August 2021
Accepted 15 September 2021
Available online 20 September 2021

Keywords:
Audio–Visual Speech Recognition
Adversarial attacks
Cross-modality
Detection network

ABSTRACT

Audio–Visual speech recognition (AVSR) is an effective way to predict text corresponding to the spoken words using both audio and face videos, even in a noisy environment. 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.

1. Introduction

Automatic Speech Recognition (ASR) models are widely used in converting of speech to text. Its application can be seen in numerous products like Amazon Alexa, Microsoft Cortana, Google Assistant and Apple Siri. The efficacy of these models drops in the presence of noise. This issue can be tackled by either mitigating the noise using speech enhancement techniques [1] or utilising visual features along with speech [2,3]. It is shown in [4] that speech to text conversion can be improved when facial expression and lip movements derived from face videos are used along with speech. These factors proliferate the research in the field of Audio–Visual Speech Recognition (AVSR) [5]. In AVSR, the features extracted from both audio speech and face videos are utilised for predicting the text corresponding to the spoken word. Some real-world applications of AVSR are: i) performing speech recognition even if one of the modalities (that is, either visual or audio) is noisy [6]; ii) speaker verification in multi-speaker scenarios [7]; iii) audio–visual speech separation [8]; iv) talking face synthesis [9]; v) aiding hearing-impaired persons by providing transcriptions [10]; vi) biometrics verification [11]; and vii) event recognition in surveillance videos [12].

The AVSR model is vulnerable to adversarial examples, like any other machine learning model. The adversarial examples are crafted by adding well-structured perturbations (or noise) to the original input such that the model makes incorrect predictions and changes are unrecognised by humans [13,14]. The vulnerability of AVSR models to adversarial examples hampers the efficacy of the applications mentioned above. Thus, there is a need to design a robust and secure AVSR model. It can be achieved by understanding the adversarial attacks and defences on AVSR models (see Fig. 1).

The AVSR models are more resilient against adversarial attacks than image classification. As opposed to images, the AVSR incorporates temporal information, which provides a defence mechanism and thereby mitigates the adversarial attacks [15]. Likewise, the existing attacks on video recognition models are not directly applicable to AVSR models because region-of-interest in the AVSR model is smaller, leading to perceptible distortions. Furthermore, fooling the AVSR model is difficult than fooling the ASR model because AVSR works on two modalities that complement each other. That is, when one modality is attacked, the other modality tries to revert the prediction to the correct label. Hence, it is challenging to perform the adversarial attack using a single modality, as in image classification, video classification and ASR model.

Adversarial examples are usually generated by backpropagating the gradients [16]. Unfortunately, the AVSR models contain non-differentiable layers, which restricts the gradient backpropagation and thereby, restricting the adversarial attack. Even if the

[*] Corresponding author.
E-mail address: deeplearning@iiti.ac.in (A.K. Gupta).

[1] Equal Contribution.

https://doi.org/10.1016/j.knosys.2021.107503
0950-7051/© 2021 Elsevier B.V. All rights reserved.

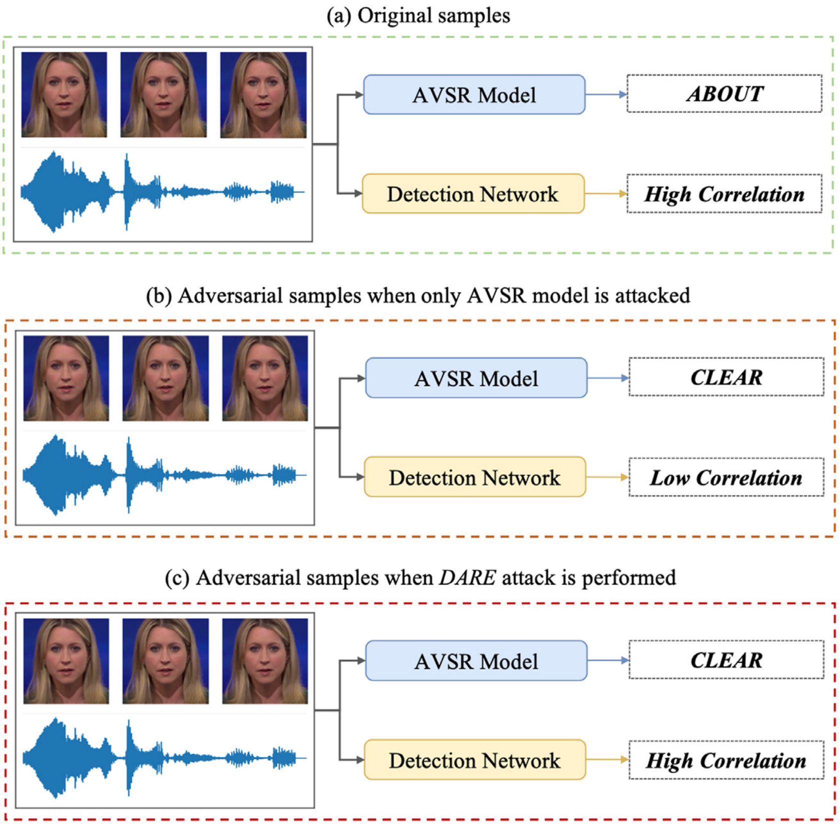

Fig. 1. Illustration of the *DARE* attack on face video and audio. (a) When passed through the AVSR model and the detection network, original inputs give the prediction "ABOUT" and a high correlation value, respectively. (b) When an attack is performed only on the AVSR model, the output of the AVSR model changes to the target "CLEAR", but the detection network outputs a low correlation, indicating an adversarial input. (c) Whereas when *DARE* is performed by attacking both the networks simultaneously, the output of AVSR is the targeted output "CLEAR" and the detection network reports a high correlation..

targeted adversarial attack is performed on the AVSR model, the adversarial example can be easily detected [17] using a detection network, SyncNet [7]. The network is designed based on the intuition that the correlation between audio and video decreases when adversarial perturbations are added to the original sample. That is, the correlation between audio and video is lower for adversarial examples than the original examples. A novel adversarial attack on the AVSR model is proposed in this paper, which handles all the aforementioned challenges. It is referred to as the Deceiving Audio–visual speech Recognition model, *DARE*. Its main contributions are:

1. To the best of our knowledge, we are the first to fool the AVSR model and detection network simultaneously to generate targeted adversarial examples. We fool the detection network to prevent it from distinguishing original and adversarial samples. For this purpose, we maintain the correlation between audio and face videos.
2. Our extensive experiments conducted using state-of-the-art AVSR model on publicly available Lip Reading in the Wild (LRW) dataset, demonstrate that attacking either audio or video modality results in perceivable distortions. Thus, the proposed attack, *DARE* achieves the desired results by adding small and imperceptible distortions to both the modalities.
3. Our proposed attack successfully circumvents popular defences with an attack success rate of 100% while maintaining imperceptibility of added perturbations.

The rest of the paper is organised into the following sections. The next section provides an outline of the existing works. The proposed attack is presented in Section 3. The experimental results are discussed in Section 4 and discussions are mentioned in Section 5. The contributions are summarised in Section 6.

2. Related work

This section provides a brief description of well-known AVSR models, adversarial attacks and the detection networks for AVSR adversarial attacks.

2.1. AVSR models

ASR models are widely used to convert spoken utterances to the corresponding transcriptions by utilising the audio modality [18]. It performs poorly when the audio is corrupted by noise [19]. Alternatively, the visual information can be used for speech recognition. For instance, the authors in [20] presented a visual speech recognition model based on LSTM using visual-only features. The efficacy of speech recognition model improves when both audio and visual modalities are simultaneously utilised, and such models are referred to as AVSR models. This observation is leveraged in [21]. It consists of two streams corresponding to audio and video. Each stream analyses the corresponding input using Bidirectional Long Short-Term Memory (Bi-LSTM). Subsequently, the resultants of both streams are consolidated to predict the transcription. Likewise, the AVSR model proposed in [5] consists of two streams corresponding to audio and video modalities. Each stream analyses the provided modality using ResNet, followed by 2-layer Bidirectional Gated Recurrent Unit (BGRU) network to model the temporal dynamics. Subsequently, the resultant of both streams are fused using another 2-layer

BGRU to predict the transcription. Several recent works [22–24] have used attention mechanism for AVSR. A study comparing the performance of sequence-to-sequence and connectionist temporal classification (CTC) prediction methods on a self-attention architecture based AVSR model is performed in [22]. It was observed that both methods performed similarly for AVSR. A hybrid architecture incorporating CTC and attention mechanism is proposed in [23]. This work is extended in [24], which employs CTC and a convolution-based attention mechanism. To the best of our knowledge, the model[2] presented in [5] is the current state-of-the-art for word recognition task on the LRW dataset.

2.2. Adversarial attacks

Adversarial examples are generated by adding imperceptible perturbations to input samples with an intention to fool the machine learning model. Based on an adversary's goal, the adversarial attacks can be categorised either as targeted or untargeted attack [25]. Targeted attacks are performed by generating an adversarial example which on classification predicts a specific target label. Whereas, in untargeted attacks, the goal is to expect any incorrect label on classification. Adversarial attacks can also be categorised according to the adversary knowledge into either white box or black box attacks [26,27]. The adversary has complete information regarding the model architecture and parameters in a white box setting. In contrast, the adversary has limited or no knowledge about the model architecture or its parameters in a black-box setting.

Several adversarial attacks have been proposed in the literature for images [28]. The simplest and fastest amongst them is the Fast Gradient Sign Method (FGSM), which uses the sign of the gradient for calculating the perturbations [16]. The gradients are calculated by taking derivatives of the loss function with respect to the input image. It gives us the contribution of each pixel in the loss function. Mathematically,

$$x^{adv} = x - \epsilon * \text{sign}(\nabla_x L(f(x), y)) \tag{1}$$

where, x is the original image, x^{adv} is the adversarial image, $f(x)$ is the original label, y is the target output label, ϵ is a step size to ensure small perturbation is added, L is loss function and ∇_x denotes derivative with respect to x. The FGSM can be improved by adding smaller perturbations at each iteration. Such an attack is called an Iterative Gradient Sign Method (IGSM) [28]. Formally, at each iteration n,

$$x_{n+1}^{adv} = x_n^{adv} - \epsilon * \text{sign}(\nabla_x L(f(x_n^{adv}), y)), \quad \text{such that } x_0^{adv} = x \tag{2}$$

where, x_n^{adv} denotes the adversarial image at nth iteration. There exist adversarial attacks on ASR, and although it is difficult to perform an adversarial attack on audio compared to images due to non-linearity in the audio domain [29]. One such adversarial attack is proposed by [29] which is an iterative and optimisation-based attack to add imperceptible perturbation to the input audio samples. It uses the following optimisation:

$$\text{minimise} \quad l(f(a+\delta), y) \quad \text{such that} \quad \|\delta\| < \epsilon \tag{3}$$

where, a is original audio, δ is added perturbation to audio sample, ϵ is used to make sure that the perturbation δ is within a small range, $f(.)$ is ASR model, l is loss function. The goal is to minimise the loss function l, which can be done when the target phrase y is predicted. The perturbations generated by attacks such as FGSM and IGSM are per-instance, that is the perturbations are specific to input for which it was generated. Whereas the universal attack generates universal perturbations which results in misclassification, when added to most of the inputs [30,31].

Similar to image and audio modalities, the AVSR models are vulnerable to adversarial attacks. For instance, an untargeted attack on the AVSR model for the publicly available and extensively utilised LRW dataset is proposed in [17]. Attacking an AVSR model is challenging due to the presence of temporal dimension, non-differentiable layers, and both the modalities complement each other. Hence, not much work has been carried out in attacking the AVSR models. Moreover, even if an adversary successfully attacks an AVSR model, it can be easily detected using a detection network [17] (refer Section 2.3).

2.3. Detection network

The audio and video streams are highly correlated for the original signals. This phenomenon is employed to improve the performance of several audio–visual tasks such as event localisation [32], action recognition [33], emotion recognition [34] and AVSR [20]. However, the correlation decreases when adversarial perturbations are added in AVSR [17]. That is, the correlation between the audio and video streams in an original sample would be higher than the adversarial sample. This observation is leveraged in [17] for detecting the adversarial examples. To the best of our knowledge, there is only one detection method for detecting the adversarial examples for AVSR [17]. It utilises the detection network[3] proposed in [7] for analysing the correlation. It provides the confidence score or correlation between the audio and video streams. It consists of two streams that take as input the MFCC features of the audio and the extracted mouth region from the face video. Subsequently, the Euclidean distance between the resultant of both streams is used to obtain the confidence score.

3. Proposed attack

This section discusses our proposed attack *DARE* to generate targeted adversarial example while remaining undetected. As discussed in Section 2.3, the detection network can detect whether the given video sample is benign or perturbed samples based on the confidence score. To prevent this detection and achieve the required target, we simultaneously fool the detection network and AVSR model. The overview of the proposed attack is shown in Fig. 2. It comprises of state-of-the-art AVSR model [5] and detection network [7]. The section first describes how to fool the individual AVSR model and detection network. Subsequently, it presents the mechanism of fooling both the AVSR model and detection network simultaneously. Eventually, the implementation details of the *DARE* attack are provided.

3.1. Attacking AVSR model

In this subsection, the adversarial attack is performed on AVSR model. The AVSR model, f takes audio, a and face videos, v as input and predict the corresponding word, $f(v, a)$ (refer Fig. 2). Our targeted attack adds the adversarial perturbations in the inputs such that the adversarial audio $\overline{a}$ sound similar to a, and the adversarial face videos $\overline{v}$ and v are visually imperceptible. However, the corresponding prediction $f(\overline{v}, \overline{a})$ corresponds to target word label, which is different than $f(v, a)$. We use cross-entropy loss for attacking the AVSR model. The cross-entropy loss can take either logits or probabilities as one of its parameters. Logits are the unnormalised probability that is given as input to the softmax function. The output of the softmax function are

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

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

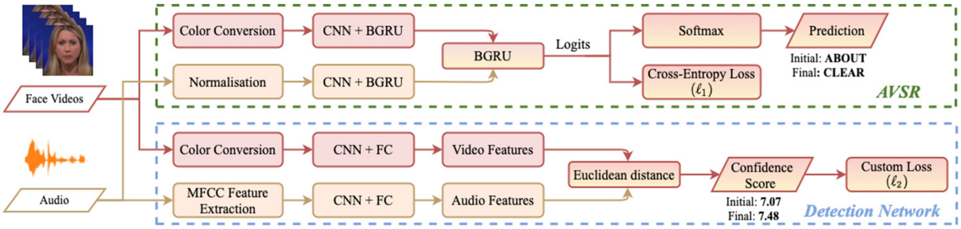

Fig. 2. Overview of the proposed *DARE* attack. It takes face videos and audio as input. The AVSR model [5] predicts the word **ABOUT** corresponding to these inputs. The detection network takes face videos and audio as inputs and computes **7.07** as a confidence score. The inputs are first given to the AVSR model to perform a targeted attack by minimising ℓ_1 loss. Subsequently, the generated adversarial sample is given to the detection network [7] to maintain the correlation between audio and face videos. These steps are repeated until the target word is achieved with a higher confidence score by minimising ℓ_2 loss. The generated adversarial samples, when given to the AVSR model and the detection network, the outputs are **CLEAR** and a confidence score of **7.48**, respectively.

probabilities of a particular word out of the set of labels. It is shown in [35] that adversarial attacks can be performed in a better way when logits are utilised instead of probability in the loss function. Our experimental results performed in Section 4 also advocated this observation. Thus, we utilise the *CrossEntropy* loss using logits. The loss function ℓ_1 can be represented as :

$$\ell_1(v, a, y) = CrossEntropy(z, y) = -\log\left(\frac{\exp(z[y])}{\sum_j \exp(z[j])}\right) \\ = -z[y] + \log\left(\sum_j \exp(z[j])\right) \tag{4}$$

where z is the logits obtained corresponding to the video v and audio a, whereas y is the target label. Kindly note that the features obtained from the audio and video streams are concatenated and passed to the BGRU layer, which provides us with the logits (refer Fig. 2). The size of the logits z, is equal to the number of classes (which is 500 for the LRW dataset).

Our targeted attack on face videos and audio samples for the AVSR model is based on the IGSM attack. In essence, the adversarial example is generated by adding well-crafted perturbations to the original input audio and face videos. The perturbations are obtained by taking derivatives of the loss function with respect to input audio and video samples. The attack is performed using:

$$v_{n+1} = v_n - \epsilon_v^a * sign(\nabla_v \ell_1(v_n, a_n, y)) \tag{5}$$

$$a_{n+1} = a_n - \epsilon_a^a * sign(\nabla_a \ell_1(v_n, a_n, y)) \tag{6}$$

where, v_n and a_n denote the adversarial face videos and audio samples at nth iterations, respectively; ℓ_1 is the cross-entropy loss; z is logits which can be obtained with the output of BGRU; y is target label; ϵ_v^A and ϵ_a^A are step sizes for video and audio modality. Kindly note that the step sizes are set to small values (refer Section 3.4 for parameter tuning) so that the adversarial examples are imperceptible. Moreover, the loss function achieves the minimum value, when the AVSR model prediction on adversarial examples, $f(\overline{v}, \overline{a})$ and the target label y are same, that is $f(\overline{v}, \overline{a}) = y$.

3.2. Attacking detection network

In this subsection, the adversarial attack is performed on the detection network. The network, s takes audio, a and face videos, v as input and predict the confidence score between the input audio and video streams (refer Fig. 2). The score provides the correlation between the audio and video streams. Thus, it is higher for the original sample as compared to the adversarial sample. It is leveraged by the detection network for distinguishing between the original and perturbed samples based on the confidence score. To prevent this detection, we fool the detection network with the aim that the confidence score of the adversarial sample should not be less than the confidence score of the corresponding original sample. To this end, we defined the custom loss for the detection network. The custom loss ℓ_2 is defined as the difference between the confidence score of original and adversarial samples. Mathematically, the custom loss is represented as:

$$\ell_2(\tau_0, \tau_a) = \max(0, \tau_0 - \tau_a) \tag{7}$$

where, τ_0 and τ_a are the confidence scores of the original and adversarial samples, respectively. The loss ℓ_2 is minimum when the confidence score of adversarial samples is greater than or equal to the original confidence score. The adversarial attack on face video and audio samples for maintaining high confidence score are as follows :

$$v_{n+1}^{adv} = v_n - \epsilon_v^S * sign(\nabla_v \ell_2(\tau_n, \tau_o)) \tag{8}$$

$$a_{n+1}^{adv} = a_n - \epsilon_a^s * sign(\nabla_a \ell_2(\tau_n, \tau_o)) \tag{9}$$

where, v_n and a_n denote the adversarial face videos and audio samples at nth iterations respectively; ℓ_2 is custom loss function; ϵ_v^S and ϵ_a^S are step sizes for face video and audio modality; τ_n is the confidence score at nth iteration. The value of step sizes ϵ_v^S and ϵ_v^A are chosen to be small value to generate imperceptible adversarial examples (refer Section 3.4 for parameter tuning).

3.3. DARE attack: Attacking AVSR and detection network

The proposed *DARE* attack is presented in this subsection. It aims to generate targeted adversarial example for the AVSR model while remaining undetected by the detection network. To this end, both the AVSR model and detection network are fooled (or attacked) simultaneously. The proposed attack is shown in Fig. 2, which consists of attacking both the AVSR model and detection network. We alternately attack the AVSR model and detection network to generate adversarial audio and face videos. More clearly, we first perform the targeted attack on the AVSR model by providing the original audio and face videos. It provides the adversarial example using Eqs. (5) and (6). The adversarial examples can be detected by a detection network. To avoid such detection, the detection network is fooled for these adversarial examples by increasing their confidence score. That is, the adversarial attack is performed on the generated examples to maximise the confidence score using Eqs. (8) and (9). Unfortunately, the resultant adversarial examples may not provide the required target label when given to the AVSR model. Thus, the AVSR model and detection network are attacked repeatedly until both are simultaneously fooled. The steps involved in the *DARE* attack are shown in Algorithm 1.

For illustration, consider Fig. 3, which shows the adversarial face video and audio example generated using our *DARE* attack. The added imperceptible perturbations are scaled up for better

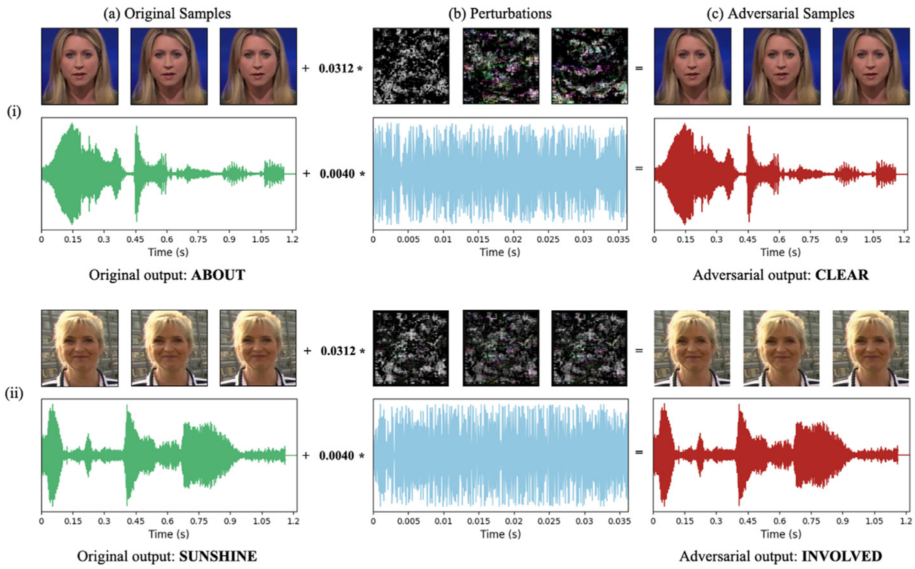

Fig. 3. Visualisation of the *DARE* attack on face video and audio in *Targeted*$_2$ setting. It depicts: (a) original samples; (b) perturbations; and (c) adversarial samples generated by *DARE* attack. The inputs for AVSR model and the detection network are lip-region and complete facial area, respectively. Hence, *DARE* attack adds perturbations in the entire facial region. The audio perturbations are plotted for a shorter duration (0.035 secs) for better visualisation. When provided to the AVSR model, the original samples provides output as "ABOUT" and "SUNSHINE", respectively. Similarly, for the adversarial samples, the outputs of AVSR model are "CLEAR" and "INVOLVED", respectively..

Algorithm 1 *DARE* Attack

```
Require: Audio $a$; face videos $v$; target label $y$; AVSR model $f(.)$;
    detection network $s(.)$; step sizes $\epsilon_v^A$ and $\epsilon_a^A$ for AVSR; and step
    sizes $\epsilon_v^S$ and $\epsilon_a^S$ for detection network
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 $s(v, a) \leq \tau_o$ do
      while $f(v, a) \neq y$ do                          ▷ Attacking AVSR
          $v = v - \epsilon_v^A * \text{sign}\left(\nabla_v \ell_1(v, a, y)\right)$     ▷ refer Eq. (5)
          $a = a - \epsilon_a^A * \text{sign}\left(\nabla_a \ell_1(v, a, y)\right)$     ▷ refer Eq. (6)
      end while
      $\tau_a = s(v, a)$
      while $\tau_a \leq \tau_o$ do              ▷ Attacking detection network
          $v = v - \epsilon_v^S * \text{sign}\left(\nabla_v \ell_2(\tau_0, \tau_a)\right)$     ▷ refer Eq. (8)
          $a = a - \epsilon_a^S * \text{sign}\left(\nabla_a \ell_2(\tau_0, \tau_a)\right)$     ▷ refer Eq. (9)
          $\tau_a = s(v, a)$
      end while
  end while
  $\overline{v} = v$
  $\overline{a} = a$
  return $(\overline{v}, \overline{a})$
```

visualisation. It is important to note that, for the first example, both the adversarial face video and audio predict the target label, "CLEAR" when provided to the AVSR model.

3.4. Implementation details

Kindly note that the AVSR model and the detection network require different ranges of input audio and face videos. For instance, the pixel intensities lie in the range of 0 to 1 for the AVSR model, while the range is 0 to 255 for the detection network. This issue is mitigated by appending a scaling layer in the preprocessing step to ensure appropriate input to the models. Likewise, the step sizes (ϵ_v^A and ϵ_a^A) are selected such that minimal changes are introduced in audio and face videos. In the case of face videos, the step size ϵ_v^A for the AVSR model is set to 0.00392 (1/255), which is the minimum possible pixel change when the range is 0 to 1. Similarly, ϵ_v^S is set to 1.0 for the detection network whose input face videos ranges from 0 to 255. However, the experiments are conducted on the audio modality for generating the imperceptible adversarial audio. The apt step size ϵ_a^A is found to be 0.00015 (5/32767) for AVSR and ϵ_a^S is 5.0 for synchronisation based detection network, respectively. Furthermore, a few non-differentiable layers are present in preprocessing, which prevents backpropagation till the original audio and face videos. It results in gradient masking, which prevents the adversarial attack [36]. This issue is tackled for face videos by replacing the non-differentiable layers with their alternative differentiable functions. The differentiable functions are provided by an open-source computer vision library Kornia [37,38]. Mainly, the image normalisation and greyscale conversion functions are replaced for the AVSR model. Similarly, the detection network requires MFCC features of the audio modality and such a feature extraction require a few non-differentiable functions. Thus, the MFCC code is rewritten for audio by following the differentiability property at each layer. It is important to note that the modified model's efficacy is the same as that of the original model.

4. Experimental results

4.1. Datasets

The publicly available dataset, LRW[4] [39] is used for conducting the experiments. It constitutes 25000 video clips in the test set. These are acquired from the broadcast content of BBC News. Each video is 1.16 s long and contains 29 frames. The dataset comprises 500 words, and the target word is present in the middle of the video. Kindly note that only those samples are used for the experimentation, which are correctly classified by the AVSR model.

[4] 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.)

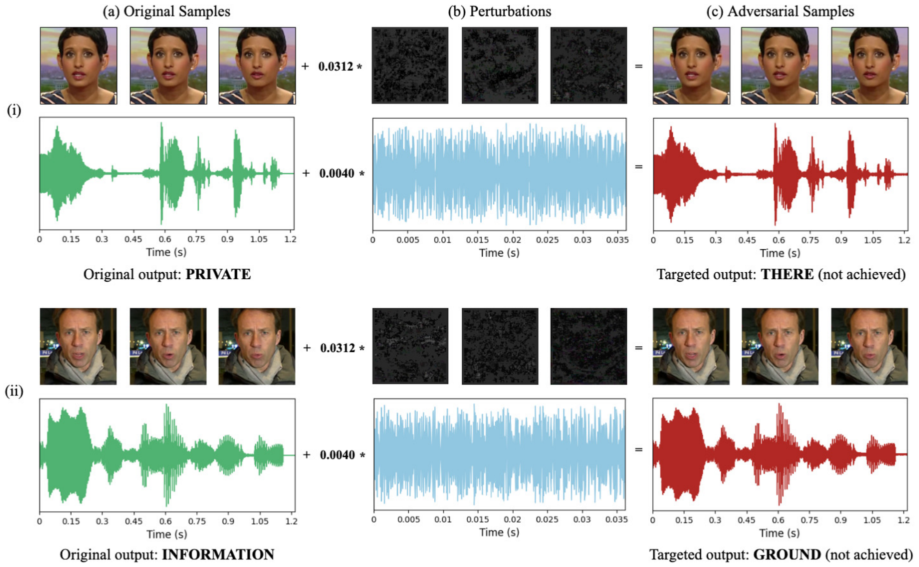

Fig. 4. Visualisation of the failure cases of the *RDARE* attack in *Targeted*$_2$ setting. It depicts: (a) original samples; (b) perturbations; and (c) adversarial samples generated. (i) The output of AVSR model for the clean input is "PRIVATE". The *RDARE* attack fails to achieve the target "THERE". (ii) The output of AVSR model for the clean input is "INFORMATION". The *RDARE* attack fails to achieve the target "GROUND"..

4.2. Experimental settings

The performance is assessed using the attack success rate, which is defined as the percentage of successfully attacked samples. Moreover, it is assessed for the video modality using the average video distortion, δ_∞. For individual samples, the video distortion is defined as the maximum intensity difference between adversarial and original face videos. Similarly, the performance of audio modality is assessed using average audio distortion, D. The distortion in audio is given by the relative loudness of the perturbation with respect to the audio. Formally, if δ and a denote the adversarial perturbation and original audio a, respectively, then the distortion $D_{a,\delta}$ is given by:

$$D_{a,\delta} = dB(\delta) - dB(a) \tag{10}$$

where, $dB(\alpha)$ represents the decibel value of the audio α, as given in [29]. The difference in Eq. (10) is always negative as the perturbation introduced is quieter than the original signal [29]. The lower value of the average distortion metric signifies quieter distortion and thereby better efficacy.

To study the efficacy of our *DARE* attack, we perform the attack in *Targeted*$_1$ and *Targeted*$_2$ settings. In the *Targeted*$_1$ setting, the attack is performed by fixing the target to the second most probable label given by the AVSR model. Thus, this attack can be easily performed by adding small adversarial perturbations in the input samples. In contrast, the labels are set to the least probable label given by the AVSR model in the *Targeted*$_2$ setting. Thus, this attack is difficult to perform and requires large adversarial perturbations.

4.3. Comparative analysis

This subsection discusses the experimental results required to understand the effectiveness of the *DARE* attack. For more rigorous analysis, the proposed *DARE* attack is compared with audio-only, video-only, combined loss attacks. In audio-only and video-only attacks, the perturbations are added to only one modality while keeping the other modality unaltered. These attacks are useful to understand the contribution of each modality in performing the attack. Similarly, we design the combined loss attack to explore an alternate way to simultaneously attack the AVSR model and detection network. The adversarial video and audio are generated in the combined loss attack using Eqs. (8) and (9), with the difference that loss $\ell_2(\tau_0, \tau_a)$ is replaced with the loss ℓ_3. The loss ℓ_3 is a combination of cross-entropy loss ℓ_1 (refer Section 3.1) and custom loss ℓ_2 (refer Section 3.2). That is,

$$\ell_3 = c_1 * \ell_1(v, a, y) + c_2 * \ell_2(\tau_o, \tau_a) \tag{11}$$

The hyperparameters c_1 and c_2 denote the contribution of the loss functions in combined loss attack. The combined loss has been tested on different values of hyperparameters, and the best values of c_1 and c_2 are found to be 1 and 9.87, respectively. Furthermore, the proposed *DARE* attack is compared with *RDARE* and *PDARE*, which are designed from *DARE* attack by restricting the video distortion of video samples to 5 and replacing logits with probability in the loss functions, respectively.

The comparative performance is shown in Table 1. All these attacks are performed by maintaining the correlation between two modalities so that the adversarial examples remain undetected by the detection network. It can be observed from the table that it is better to perform attacks on both audio and video modalities than on a single modality. It is because if adversarial perturbations can be added in only one modality (as in audio-only and video-only attack), then large perturbations are required. Moreover, it can be observed from the table that the *DARE* attack performs better than the combined loss attack, as *DARE* attack can successfully attack the AVSR model by introducing less distortions in both the modalities. Also, it can be seen that *DARE* significantly outperforms *PDARE*, which indicates that there is performance degradation when logits are replaced by probabilities in the loss function. This observation is in agreement with the observation in [35]. The *PDARE* attack is performed by restricting the video distortion δ_∞ to 20 pixels and audio distortion D to −30 dB. This attack introduces higher distortions in audio and face videos, due to which the attack success rate is very low for the *Targeted*$_2$ label.

The proposed *DARE* attack achieves a 100% targeted attack success rate by introducing less average audio and video distortions. Usually, the distortion in the video is set to 5 pixels, and in such a case, it is not possible to attack the AVSR model for all the target labels. Hence, the attack success rate is less than 100% in *RDARE*. This effect is prominent when *Targeted*$_2$ settings are used because it is difficult to perform and require

Table 1
Comparative performance analysis of the proposed *DARE* attack.

Attack types	Targeted$_1^*$			Targeted$_2^*$		
	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
RDARE^1	99.19	2.68	−55.27	20.45	4.82	−38.93
PDARE^2	100.0	2.83	−52.27	35.23	18.16	−33.71
DARE	**100.0**	**2.74**	**−57.42**	**100.0**	**8.73**	**−46.18**

*: The target label is set to the second-most and least probable label in *Targeted*$_1$ and *Targeted*$_2$ setting, respectively.
$^+$: The lower value of δ_∞ or D indicate better imperceptibility and thereby better performance.
$^-$: Audio-only attacks are performed by making changes in audio and keeping face videos unaltered.
$^{--}$: Video-only attacks are performed by making changes in face videos and keeping audio unaltered.
^{1}: *RDARE* is designed from *DARE* by restricting δ_∞ to 5.
^{2}: *PDARE* is designed from *DARE* by replacing logits with probability in the loss functions and by restricting D to −30 dB and δ_∞ to 20.
Note : The attacks are performed on only those audio–visual samples which are correctly classified by the AVSR model.

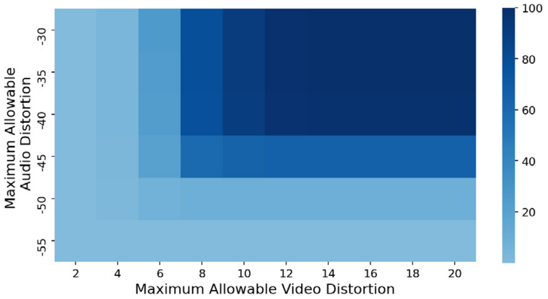

Fig. 5. Heatmap of attack success rate in *Targeted*$_2$ setting for *DARE* attack. The x and y-axis denote the maximum allowable distortion in video (δ_∞) and audio (D), respectively.

large perturbations. Examples of the failed examples of *RDARE* attack is illustrated in Fig. 4. Kindly note that it is possible to achieve a 100% attack success rate by reducing the step-sizes like, setting the step-size less than 1 in video modality. However, this paper avoids the mechanism because it will generate those adversarial examples that cannot be saved and reused back to fool the AVSR models later due to quantisation error. Consider Fig. 5 for better understanding the efficacy of the proposed *DARE* attack in *Targeted*$_2$ setting. It depicts the attack success rate of *DARE* by restricting the maximum values of δ_∞ and D, using a heatmap.

We also study the applicability of universal adversarial attacks on AVSR. We extend the image classification attack proposed in [30] for attacking the AVSR and detection networks simultaneously. Kindly note that we add the perturbations only in the video modality while keeping the audio modality unaltered. We perform our experiments on the LRW dataset. We use the videos from the train set to generate a universal perturbation, V_p. The perturbation is then added to the videos of the test set for the evaluation. Since the perturbation added is the same for every input video in the test set, we have used a different evaluation metric, which is Δ_∞, the highest intensity value of the perturbation V_p. Furthermore, due to a high computational cost, we use 7000 videos from the training set to generate the perturbation and 5000 videos from the testing set for evaluation. It was observed that the value of Δ_∞ was 42.63 in our case, and hence the perturbations added to the video modality were easily perceivable. We observed that the adversarial samples generated by the attack were successfully misclassified by the AVSR model. However, in our case, an attack is considered successful when the adversarial video is misclassified by the AVSR while remaining undetected by the detection network (refer sec. 3.3). In our case, the synchronisation-based detection network could easily detect the adversarial samples because the correlation between the audio and video modalities decreases when large perturbations are added (which is Δ_∞ = 42.63 in our case). Hence, we get a low success rate of 15.47% when we apply the universal attack [30] on AVSR.

4.4. Impact of defences on DARE attack

In the literature, it is shown that several powerful attacks can be easily mitigated by adversarial defences and thus, these attacks are of limited applicability [40]. Hence, we analyse the impact of the popular defences of audio and video modalities on our *DARE* attack for a more thorough understanding. In essence, the performance of *DARE* attack is analysed after the popular input transformation defences are applied to mitigate the attack. To this end, the following three popular audio-based defences are used: (i) MP3 Compression, which takes audio and compresses it at a constant bit rate of 48kbps [41]; (ii) Re-sampling, where given audio is re-sampled to 8 kHz and again re-sample the audio back to 16 kHz [41]; and (iii) White noise addition, where white noise is added to the audio at a signal to noise ratio (SNR) of 60 dB [41]. The adversarial attacks and defences are not extensively studied in video modality. Thus, the following four popular image-based defences are used for the video modality: i) Bit Reduction, where bit frames are reduced from 8 to 5 bits, as recommended in [42]; ii) JPEG-Compression, where the face videos are compressed at quality level 75 (out of 100) [42]; iii) Box Blur, where each pixel of an image is replaced by the average of its neighbouring pixels in 3 × 3 neighbourhood [37]; and iv) Median Blur, where the central pixel is replaced with the median of the neighbouring pixels in 3 × 3 neighbourhood [43]. Table 2 provides the experimental results, obtained using *DARE* attack when different possible combinations of audio and video defences are applied. Some of the defences are non-differentiable, so the Backward Pass Differentiable Approximation (BPDA) is used for the analysis, as suggested by [36]. The BPDA is used to approximate the derivatives of non-differentiable functions. It can be observed from the table that the *DARE* attack can successfully circumvent the popular input transformation defences. Also, the AVSR model accuracy decreases slightly when the defences are applied. Moreover, there is an increase in average video and audio distortions when the input transformation defences are utilised. It happens because the added defences revoke some adversarial

Table 2
Performance of the *DARE* attack when popular defences are utilised to stop the attack.

Defence used	Model$^{\#}$ accuracy (in %)	Targeted$_1^*$		Targeted$_2^*$	
		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 and least probable label in *Targeted*$_1$ and *Targeted*$_2$ setting, respectively.
$^+$: The lower value of δ_∞ or D indicate better imperceptibility and thereby better performance.
$^\#$: The percentage of original audio–visual samples that are correctly classified by AVSR model.
Note: We perform the attacks on only those audio–visual samples which are correctly classified by the AVSR model.

perturbations. Further, it can be seen that the proposed *DARE* attack can easily handle the Box Blur and Re-sampling audio defences for video and audio modality, respectively. However, the most challenging defence for the AVSR model is applying JPEG-Compression and MP3-Compression in video and audio, respectively.

A recent work on audio–visual video parsing tasks leverages the strong correspondence within the audio and visual streams of a video to attain state-of-the-art performance [44]. It tackles the issue of audio–visual asynchrony present in the video due to off-the-screen sounds. To this end, the work proposes to exchange the audio and visual track of video in the training set with other non-related videos. We study the idea of swapping the audio of one video with another as a potential defence against the proposed attack. For a given video, we exchange the audio stream with a random audio of another video from the dataset and keep the video stream unaltered. We observed a significant drop of 86.92% in the accuracy of the AVSR model when we apply this method as a defence. The reason for such behaviour can be linked to the strong correlation between the audio and visual modalities in the AVSR model [17]. We also observed that the audio stream plays a more dominant role in determining the predictions of such video samples. For most of such videos, we got the output label corresponding to the audio stream; for some, the output corresponding to the video stream; and for the remaining, we got the output corresponding to neither. We also studied the behaviour of the detection network in the presence of such video samples. We noted a drop in the confidence value, even for the non-attacked samples for which the audio was swapped. The reason for this behaviour is the fact that audio and visual modalities play a crucial role in AVSR [17].

5. Discussion

The AVSR attack proposed in [17] utilises the IGSM. It achieves the video and audio distortion of 1.99 and −30.54 dB, respectively, in *Targeted*$_1$ setting. Likewise, it achieves the video and audio distortion of 13.43 and −15.38 dB, respectively in *Targeted*$_2$ setting. To compare the results with the attack in [17], AAVSR attack is designed by avoiding the attack on detection network in *DARE*. Thus, AAVSR attack performs the attack only on the AVSR model. In *Targeted*$_1$ setting, the average video and audio distortions are 1.87 and −60.64 dB respectively. Similarly for *Targeted*$_2$ setting, the average video and audio distortions are 8.86 and −47.16 dB respectively. It can be observed that the AAVSR attack outperforms the attack proposed in [17]. It is because the AAVSR loss function makes use of logits instead of probabilities. Moreover, the step size for audio in AAVSR is small, which allows finer perturbation at each iteration. However, both AAVSR and attack proposed in [17], can be easily detected by a detection network. Even though the distortion introduced in *DARE* is comparatively larger than that of the AAVSR attacks, we advocate using *DARE* as the perturbations added by *DARE* remains undetectable.

Our proposed *DARE* attack is a generic attack that will work for any other AVSR models and detection networks, with slight modifications. The AVSR models can be easily fooled using gradient-based attacks. However, if gradient backpropagation is obstructed by non-differentiable layers, then these layers need to be replaced by differentiable layers, as discussed in Section 3.4. Similarly, any detection network will output the confidence score; thus, our custom based loss function will be directly utilised to maintain the correlation between the two modalities. To the best of our knowledge, there is no other pre-trained model available for AVSR on the publicly available dataset. Hence, the experiments are conducted on the state-of-the-art AVSR model on the LRW dataset.

To understand the efficacy of *DARE* attack on other audio–visual tasks, we perform the attack on audio–visual action recognition model. The adversarial videos generated by the *DARE* attack can successfully fool the classification network while remaining undetected by the synchronisation-based detection network. Hence, we need both the classification and synchronisation-based detection networks to study the efficacy of the proposed attack. To this end, we use a fine-tuned MC3-VGG architecture based audio–visual action recognition model[5] and MC2-VGG architecture based detection model[6] for performing the attack [33]. The fine-tuning and attack is performed on a subset of the publicly available Kinetics [45] dataset. When we perform the *DARE* attack in *Targeted*$_1$ and *Targeted*$_2$ settings, we achieve 100% attack success rate. The average video distortion δ_∞ in *Targeted*$_1$ and *Targeted*$_2$ settings are 2.97 and 7.87 respectively. Moreover, we observe that the video distortions introduced for the action recognition task are slightly lower compared to that of AVSR (refer Table 1). This behaviour can be attributed to the smaller region-of-interest in AVSR compared to the action recognition model. For action recognition task, the average audio distortions *D*, in *Targeted*$_1$ and *Targeted*$_2$ settings were −56.42 dB and −49.21 dB respectively. These distortions are comparable to the audio distortions of AVSR, because of similar audio modalities.

6. Conclusion and future work

AVSR models are utilised widely in many applications, and adversarial attacks on these AVSR models will cause profound implications. A thorough understanding of such attacks is necessary for developing robust AVSR models. Attacking an AVSR model is challenging because both audio and visual modalities complement each other. Moreover, these attacks can be easily identified using detection networks. Thus, a first-ever attack, *DARE* attack, has been proposed in this paper. It has successfully attacked the well known AVSR model by adding imperceptible perturbations and simultaneously remaining undetected by the detection network. Our extensive experiments have shown that attacking either audio or video modality results in perceivable distortions. Thus, the proposed *DARE* attack has added imperceptible perturbations in both the modalities. Moreover, the experiments have demonstrated that *DARE* attack successfully attains any target even when combinations of popular image and audio based defences are utilised to mitigate the attack. The success of adversarial attacks stems from the fact that while building AVSR models, the main focus is to improve its accuracy instead of making it robust. We strongly believe that our work will create a new research direction to understand and design reliable and robust AVSR models without sacrificing accuracy. We will extend this work by further exploring the possibility of building better defences and stringent detection algorithms to mitigate such types of attacks. Furthermore, in future, we will study the applicability of the adversarial attacks and defenses in other extensively explored audio–visual tasks, thereby making them more robust. Some such tasks are audio–visual video parsing [44], audio–visual event localisation, and cross-modality localisation [32].

CRediT authorship contribution statement

Saumya Mishra: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Anup Kumar Gupta:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Puneet Gupta:** Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration.

[5] Pre-trained model available at: https://vlg.cs.dartmouth.edu/projects/avts/model_mc3_as.pt.

[6] Pre-trained model available at: https://vlg.cs.dartmouth.edu/projects/avts/model_mc2_as.pt.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We want to pay our gratitude to the respective authors for providing code and pre-trained models. We are also grateful to BBC Research for providing us with the LRW dataset. We are also thankful to the anonymous reviewers for their valuable suggestions to improve the quality of the paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] C. Donahue, B. Li, 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, http://dx.doi.org/10.1109/ICASSP.2018.8462581.

[2] G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-visual automatic speech recognition: An overview, Issues Vis. Audio-Vis. Speech Process. 22 (2004) 23.

[3] A.M. Barbancho, L.J. Tardón, J. López-Carrasco, J. Eggink, I. Barbancho, Automatic classification of personal video recordings based on audiovisual features, Knowl.-Based Syst. 89 (2015) 218–227, http://dx.doi.org/10.1016/j.knosys.2015.07.005, URL https://www.sciencedirect.com/science/article/pii/S0950705115002579.

[4] W.H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am. 26 (2) (1954) 212–215, http://dx.doi.org/10.1121/1.1907309.

[5] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic, End-to-end audiovisual speech recognition, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6548–6552, http://dx.doi.org/10.1109/ICASSP.2018.8461326.

[6] D. Stewart, R. Seymour, A. Pass, J. Ming, Robust audio-visual speech recognition under noisy audio-video conditions, IEEE Trans. Cybern. 44 (2) (2014) 175–184, http://dx.doi.org/10.1109/TCYB.2013.2250954.

[7] J. Chung, 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, http://dx.doi.org/10.1007/978-3-319-54427-4_19.

[8] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.T. Freeman, M. Rubinstein, Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation, ACM Trans. Graph. 37 (4) (2018) 112:1–112:11, http://dx.doi.org/10.1145/3197517.3201357.

[9] J.S. Chung, A. Jamaludin, A. Zisserman, You said that? in: British Machine Vision Conference (BMVC), BMVA Press, 2017, pp. 109.1–109.12, http://dx.doi.org/10.5244/C.31.109.

[10] P. Borde, A. Varpe, R. Manza, P. Yannawar, Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition, Int. J. Speech Technol. 18 (2) (2015) 167–175, http://dx.doi.org/10.1007/s10772-014-9257-1.

[11] P.S. Aleksic, A.K. Katsaggelos, Audio-visual biometrics, Proc. IEEE 94 (11) (2006) 2025–2044, http://dx.doi.org/10.1109/JPROC.2006.886017.

[12] M. Cristani, M. Bicego, V. Murino, Audio-visual event recognition in surveillance video sequences, IEEE Trans. Multimed. 9 (2) (2007) 257–267, http://dx.doi.org/10.1109/TMM.2006.886263.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations (ICLR), 2014, URL http://arxiv.org/abs/1312.6199.

[14] P. Gupta, E. Rahtu, Mlattack: Fooling semantic segmentation networks by multi-layer attacks, in: German Conference on Pattern Recognition (GCPR), Springer, 2019, pp. 401–413, http://dx.doi.org/10.1007/978-3-030-33676-9_28.

[15] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy (SP), IEEE Computer Society, 2017, pp. 39–57, http://dx.doi.org/10.1109/SP.2017.49.

[16] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations (ICLR), 2015, URL http://arxiv.org/abs/1412.6572.

[17] P. Ma, S. Petridis, M. Pantic, Detecting adversarial attacks on audio-visual speech recognition, 2019, arXiv preprint arXiv:1912.08639.

[18] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: International Conference on Machine Learning (ICML), PMLR, 2014, pp. 1764–1772, URL http://proceedings.mlr.press/v32/graves14.html.

[19] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (9) (2003) 1306–1326, http://dx.doi.org/10.1109/JPROC.2003.817150.

[20] S. Petridis, Z. Li, 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, http://dx.doi.org/10.1109/ICASSP.2017.7952625.

[21] S. Petridis, Y. Wang, Z. Li, M. Pantic, End-to-end audiovisual fusion with LSTMs, in: Auditory-Visual Speech Processing (AVSP), ISCA, 2017, pp. 36–40, http://dx.doi.org/10.21437/AVSP.2017-8.

[22] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2018) http://dx.doi.org/10.1109/TPAMI.2018.2889052.

[23] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, M. Pantic, Audio-visual speech recognition with a hybrid ctc/attention architecture, in: IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 513–520, http://dx.doi.org/10.1109/SLT.2018.8639643.

[24] P. Ma, S. Petridis, M. Pantic, End-to-end audio-visual speech recognition with conformers, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 7613–7617, http://dx.doi.org/10.1109/ICASSP39728.2021.9414567.

[25] P. Rathore, A. Basak, S.H. Nistala, 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, http://dx.doi.org/10.1109/IJCNN48605.2020.9207272.

[26] B. Alshemali, J. Kalita, Improving the reliability of deep neural networks in NLP: a review, Knowl.-Based Syst. 191 (2020) 105210, http://dx.doi.org/10.1016/j.knosys.2019.105210, URL https://www.sciencedirect.com/science/article/pii/S0950705119305428.

[27] K. Ding, X. Liu, W. Niu, T. Hu, Y. Wang, X. Zhang, A low-query black-box adversarial attack based on transferability, Knowl.-Based Syst. 226 (2021) 107102, http://dx.doi.org/10.1016/j.knosys.2021.107102, URL https://www.sciencedirect.com/science/article/pii/S0950705121003658.

[28] A. Kurakin, I.J. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: International Conference on Learning Representations (ICLR), 2017, URL https://openreview.net/forum?id=HJGU3Rodl.

[29] N. Carlini, D. Wagner, Audio adversarial examples: Targeted attacks on speech-to-text, in: IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 1–7, http://dx.doi.org/10.1109/SPW.2018.00009.

[30] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: Conference on Computer Vision and Pattern Recognition CVPR, IEEE Computer Society, 2017, pp. 86–94, http://dx.doi.org/10.1109/CVPR.2017.17.

[31] L. Wang, X. Chen, R. Tang, Y. Yue, Y. Zhu, X. Zeng, W. Wang, Improving adversarial robustness of deep neural networks by using semantic information, Knowl.-Based Syst. 226 (2021) 107141, http://dx.doi.org/10.1016/j.knosys.2021.107141, URL https://www.sciencedirect.com/science/article/pii/S0950705121004044.

[32] Y. Wu, L. Zhu, Y. Yan, Y. Yang, Dual attention matching for audio-visual event localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Wu_Dual_Attention_Matching_for_Audio-Visual_Event_Localization_ICCV_2019_paper.pdf.

[33] B. Korbar, D. Tran, L. Torresani, Cooperative learning of audio and video models from self-supervised synchronization, in: Advances in Neural Information Processing Systems (NIPS), 2018, pp. 7773–7784, URL https://proceedings.neurips.cc/paper/2018/file/c4616f5a24a66668f11ca4fa80525dc4-Paper.pdf.

[34] N. Hajarolasvadi, H. Demirel, Deep emotion recognition based on audio–visual correlation, IET Comput. Vis. 14 (7) (2020) 517–527, http://dx.doi.org/10.1049/iet-cvi.2020.0013.

[35] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, C.-J. 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, http://dx.doi.org/10.18653/v1/P18-1241.

[36] A. Athalye, N. Carlini, D.A. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in: International Conference on Machine Learning (ICML), Vol. 80, PMLR, 2018, pp. 274–283, URL https://arxiv.org/abs/1802.00420.

[37] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, G. Bradski, Kornia: an open source differentiable computer vision library for pytorch, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3674–3683, http://dx.doi.org/10.1109/WACV45572.2020.9093363.

[38] E. Riba, M. Fathollahi, W. Chaney, E. Rublee, G. Bradski, torchgeometry: when PyTorch meets geometry, in: PyTorch Developer Conference, 2018.

[39] J.S. Chung, A. Zisserman, Lip reading in the wild, in: Asian Conference on Computer Vision (ACCV), Springer, 2016, pp. 87–103, http://dx.doi.org/10.1007/978-3-319-54184-6_6.

[40] P. Gupta, E. Rahtu, Ciidefence: defeating adversarial attacks by fusing class-specific image inpainting and image denoising, in: International Conference on Computer Vision (ICCV), 2019, pp. 6708–6717, http://dx.doi.org/10.1109/ICCV.2019.00681.

[41] V. Subramanian, E. Benetos, M.B. Sandler, Robustness of adversarial attacks in sound event classification, in: Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2019, pp. 239–243, http://dx.doi.org/10.33682/sp9n-qk06.

[42] C. Guo, M. Rana, M. Cissé, L. van der Maaten, Countering adversarial images using input transformations, in: International Conference on Learning Representations (ICLR), 2018, URL https://openreview.net/forum?id=SyJ7ClWCb.

[43] W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, in: Annual Network and Distributed System Security Symposium (NDSS), The Internet Society, 2018, URL https://arxiv.org/pdf/1704.01155.pdf.

[44] Y. Wu, Y. Yang, Exploring heterogeneous clues for weakly-supervised audio-visual video parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1326–1335, URL https://openaccess.thecvf.com/content/CVPR2021/papers/Wu_Exploring_Heterogeneous_Clues_for_Weakly-Supervised_Audio-Visual_Video_Parsing_CVPR_2021_paper.pdf.

[45] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017, CoRR arXiv:1705.06950.