Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

TC 11 Briefing Papers

On the human evaluation of universal audio adversarial perturbations

*Jon Vadillo**, *Roberto Santana*

Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian 20018, Spain

ARTICLE INFO

Article history:
Received 29 May 2021
Revised 3 September 2021
Accepted 1 October 2021
Available online 8 October 2021

Keywords:
Adversarial examples
Deep neural networks
Speech command classification
Robust speech recognition
Human perception

ABSTRACT

Human-machine interaction is increasingly dependent on speech communication, mainly due to the remarkable performance of Machine Learning models in speech recognition tasks. However, these models can be fooled by adversarial examples, which are inputs intentionally perturbed to produce a wrong prediction without the changes being noticeable to humans. While much research has focused on developing new techniques to generate adversarial perturbations, less attention has been given to aspects that determine whether and how the perturbations are noticed by humans. This question is relevant since high fooling rates of proposed adversarial perturbation strategies are only valuable if the perturbations are not detectable. In this paper we investigate to which extent the distortion metrics proposed in the literature for audio adversarial examples, and which are commonly applied to evaluate the effectiveness of methods for generating these attacks, are a reliable measure of the human perception of the perturbations. Using an analytical framework, and an experiment in which 36 subjects evaluate audio adversarial examples according to different factors, we demonstrate that the metrics employed by convention are not a reliable measure of the perceptual similarity of adversarial examples in the audio domain.

1. Introduction

Human-computer interaction increasingly relies on Machine Learning (ML) models such as Deep Neural Networks (DNNs) trained from, usually large, datasets (Fang et al., 2018; Gao et al., 2019; Hassan et al., 2018; Nunez et al., 2018). The ubiquitous applications of DNNs in security-critical tasks, such as face identity recognition (Parkhi et al., 2015; Sun et al., 2014), speaker verification (Heigold et al., 2016; Huang and Pun, 2020; Snyder et al., 2017), voice controlled systems (Boles and Rad, 2017; Feng et al., 2017; Gong and Poellabauer, 2018) or signal forensics (Bayar and Stamm, 2018; Athulya, Sathidevi, et al., 2017; Bayar and Stamm, 2018; Zeng, Zeng, Qiu, 2017) require a high reliability on these computational models. However, it has been demonstrated that such models can be fooled by perturbing an input sample with malicious and quasi-imperceptible perturbations. These attacks are known in the literature as adversarial examples (Goodfellow et al., 2014; Szegedy et al., 2014). Due to the fact that these attacks are de-

* Corresponding author.
E-mail addresses: jon.vadillo@ehu.eus (J. Vadillo), roberto.santana@ehu.eus (R. Santana).
https://doi.org/10.1016/j.cose.2021.102495

signed to be hardly detectable, they suppose a serious concern regarding the reliable application of DNNs in adversarial scenarios.

The study of adversarial examples has focused primarily on image domain and computer vision tasks (Akhtar and Mian, 2018), whereas domains such as text or audio have received much less attention. In fact, such domains imply additional challenges and difficulties. One of the evident differences between domains is the way in which the information is represented, and, therefore, the way in which adversarial perturbations are measured, bounded and perceived by human subjects.

In the image domain, ℓ_p norms are mainly used as a basis to measure the distortion between the original signal and the adversarial example. However, recent works have pointed out that such metrics do not always properly represent the perceptual distortion introduced by adversarial perturbations (Dukler et al., 2019; Fezza, Bakhti, Hamidouche, Déforges, 2019; Jordan, Manoj, Goel, Dimakis, 2019). Although in some works in the audio domain ℓ_p norms are also used during the generation of adversarial examples to limit the amount of perturbation (Alzantot et al., 2018a; Gong and Poellabauer, 2017), more representative metrics are usually employed for acoustic signals, such as signal-to-noise ratio (SNR) (Du et al., 2020; Yakura and Sakuma, 2019) or Sound Pressure Level (SPL) (Abdoli et al., 2019; Roy et al., 2017; Zhang et al., 2017a). These metrics are computed in decibels (dB), which is a standard scale employed for acoustic signals. However, even for such metrics, measuring the perceptual distortion of the attacks is not straightforward, as other characteristics such as time-frequency properties (Bosi and Goldberg, 2012) have a high influence. In text problems, the difficulty of characterizing the perceptual distortion is even greater, due to the fact that every change is inevitably noticeable, and therefore, the aim is to produce semantically and syntactically similar adversarial examples (Alzantot et al., 2018b).

In this paper we focus on the human evaluation of adversarial examples in the audio domain. A more comprehensive approach to evaluating adversarial distortions can serve to better understand the risks of adversarial attacks in the audio domain. For instance, the development of adversarial defenses or secure human machine interaction systems can focus on the more effective, unnoticeable, attacks.

In particular, we focus on the hypothesis that the suitability of the approaches used in the literature to measure the amount of distortion in speech signals is questionable, and that different alternatives to evaluate the distortion in a more rigorous way should be employed, such as considering different metrics in different specific parts of the signals (Vadillo and Santana, 2019). Therefore, the goal of this study is to perform an analysis of the human perception of audio adversarial perturbations according to different factors, to test these hypotheses, and based on these results, to determine whether the similarity-metrics employed in the literature are suitable to model such subjective criterion.

The main contributions of this work are the following:

- We propose a novel experimental design to evaluate the human perception of audio adversarial examples for speech recognition tasks, according to different factors.
- We compared different distortion metrics in order to assess their suitability to provide a realistic measurement of the distortion for voice signals.
- We provide evidence that standard distortion metrics employed in previous works are not a reliable measure of the perceptual distortion of audio adversarial examples in this domain, showing that more specific metrics are required in order to achieve more realistic results.

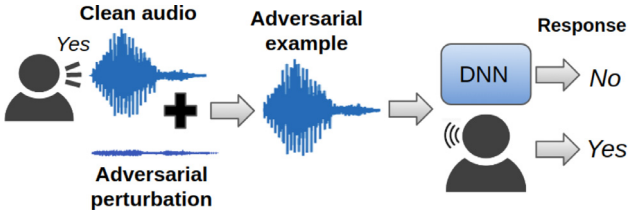

Fig. 1 – Illustration of an adversarial attack, in which an adversarial perturbation is added to a clean audio waveform, forming an *adversarial example* which is misclassified by a target DNN model, while not altering the human perception of the audio.

The remainder of the paper is organized as follows: In the following section we introduce the main concepts related to adversarial examples and review previous approaches to evaluate the distortion produced by adversarial perturbations in the audio domain. This section also highlights a number of research questions related to the evaluation of audio distortion that have not been previously addressed. Section 3 describes the selected task, target model and dataset, as well as the particular method employed for generating adversarial perturbations in the audio domain. Section 4 presents a preliminary evaluation of the adversarial perturbations according to the metrics proposed in the literature. In Section 5, we present the design of an experiment to find answers to some of the issues involved in the perceptual evaluation of the perturbations. The results of the experiment in which 36 human subjects evaluate different aspects of the adversarial perturbations are also presented and discussed. Section 6 concludes the paper and identifies lines for future research.

2. Related work

The existence of adversarial examples which are able to fool DNNs have been reported for many different audio related tasks, such as automatic speech recognition (Alzantot et al., 2018a; Carlini and Wagner, 2018; Neekhara et al., 2019), music content analysis (Kereliuk et al., 2015) or sound classification (Abdoli et al., 2019). A common adversarial attack scheme is represented in Fig. 1. Note that it is assumed that an adversary can feed the perturbed signal directly into the model. Even if this is a common assumption, some works have demonstrated that such attacks can be designed to work in the physical world (Carlini et al., 2016; Qin et al., 2019; Yakura and Sakuma, 2019; Yuan et al., 2018).

2.1. Adversarial example: Formal description

Let $f(x)$ be a classification model $f : \mathbb{X} \rightarrow \mathbb{Y}$, which classifies an input x from the input space $\mathbb{X} \subseteq \mathbb{R}^d$ as one of the classes represented in $\mathbb{Y} = \{y_1, \ldots, y_k\}$. An adversarial example x' is defined as $x' = x + v$, where $v \in \mathbb{R}^d$ represents the adversarial perturbation capable of producing a misclassification of f for the (correctly classified) input x: $f(x') \neq f(x)$. A necessary requirement for an adversarial attack is that the perturbation should be *imperceptible*, and therefore, the goal is to minimize the distortion introduced by v as much as possible, according to a suitable distortion metric $\varphi(x, x') \rightarrow \mathbb{R}$

Depending on the objective of the attack, adversarial examples can be categorized in different ways. First of all, a *targeted* adversarial example consists of a perturbed sample $x' = x + v$ which satisfies $f(x') = y_t$, where y_t represents the target (incorrect) label that we want to be produced by the model. In contrast, an *untargeted* adversarial example only requires the output label to be incorrect $f(x') \neq f(x)$, without any additional regard about the output class assigned to x'.

Furthermore, depending on the scope of the adversarial perturbation v, we can differentiate between *individual* or *universal* adversarial perturbations. In the first case, the perturbation is crafted specifically to be applied to one particular input x. Therefore, it is not expected that the same perturbation will be able to fool the model for a different sample. In the second case, universal adversarial perturbations are *input agnostic* perturbations able to fool the model independently of the input. Universal perturbations allow adversarial attacks to be produced in scenarios where individual perturbations are impractical (for instance, scenarios requiring a fast or real-time computation of adversarial examples), or performing a high number of attacks more efficiently, avoiding having to generate a new (individual) perturbation for each new input.

In Vadillo and Santana (2019), different levels of universality are proposed, depending on the number of classes for which it is expected to work. The first universality level comprises *single-class* universal perturbations that are conceived to fool the target model only for inputs of one particular class (Gupta et al., 2019; Vadillo and Santana, 2019). We will focus on *single-class* universal perturbations, although our findings regarding the weaknesses and gaps in the evaluation of adversarial perturbations are not restricted to this universality level.

2.2. Methods for assessing audio adversarial perturbations

In this section we review the strategies employed by previous works in order to verify that audio perturbations are not detectable by humans. Even if an essential requirement for adversarial perturbations to suppose a real threat is that they must be imperceptible, a good specification of such (mainly subjective) a constraint is not straightforward, and, indeed, is not well established yet.

Furthermore, even if the analysis is constrained to the audio domain, the understanding and definition of what can make a sample natural is very related to the ML task that is being solved by the model (e.g., it might be harder to categorize a music tune as "unnatural" than a spoken command). With a large variety of ML tasks related to the analysis of acoustic signals (e.g., speech recognition, music content analysis or ambient sound classification), each of them may require, therefore, a different criterion to assess the distortion of the adversarial examples according to human perception. Although a number of strategies have been proposed in these domains (Carlini and Wagner, 2018; Kereliuk et al., 2015; Roy et al., 2017; Schönherr et al., 2018; Zhang et al., 2017a), we focus our review of related work on those suitable for spoken commands. Among these strategies are:

- Thresholding the perturbation amount
- Models of human perception and hearing system
- Human evaluation

2.2.1. Thresholding the perturbation amount

The methods discussed in this section rely on limiting or measuring the perturbation amount that is added to the original input, according to a distortion metric, to ensure that the perturbations are imperceptible or *quasi-imperceptible*, or that the distortion levels are below a maximum acceptable threshold.

In Alzantot et al. (2018a), the perturbation applied to spoken commands is restricted to the 8 least-significant-bits of a subset of samples in a 16 bits-per-sample audio file. Similarly, in Gong et al. (2019); Gong and Poellabauer (2017), the effectiveness of the proposed attacks for speech paralinguistic and speech classification tasks is measured for different perturbation amounts under ℓ_p norms. The restrictions applied in these cases guarantee that the maximum change applicable to each signal is constrained. However, such thresholds are not representative for acoustic signals, as they do not guarantee a low perceptual distortion on audio attacks.

In Abdoli et al. (2019); Carlini and Wagner (2018); Neekhara et al. (2019); Yang et al. (2018), in which audio adversarial perturbations for speech recognition models are addressed, the relative loudness of the adversarial perturbation v with respect to the original signal x is measured in Decibels (dB), which is a more representative metric for acoustic signals:

$$dB_{x,max}(v) = dB_{max}(v) - dB_{max}(x), \tag{1}$$

where

$$dB_{max}(x) = \max_i \; 20 \log_{10}(|x_i|) \tag{2}$$

In Abdoli et al. (2019); Du et al. (2020), the signal to noise ratio (SNR) is used to measure the relative distortion of adversarial perturbations for speech recognition models, computed as:

$$SNR(x, v) = 10 \log_{10} \frac{P(x)}{P(v)}, \tag{3}$$

where $P(x)$ and $P(v)$ represent the power of the clean signal x and the perturbation v, respectively. The SNR has been used in other works on audio adversarial examples (Abdoli et al., 2019; Carlini et al., 2016; Kereliuk et al., 2015; Yakura and Sakuma, 2019; Yuan et al., 2018). However, these works are not based on speech signals, as their approaches rely on data with very different characteristics, such as urban sound classification, music content analysis or the injection of malicious commands

into songs. Therefore, the results are not directly comparable to spoken speech recognition, the task addressed in this paper.

2.2.2. *Models of human perception and hearing system*

The human hearing system is able to identify sounds in a range from 20Hz to 20kHz, so that perturbations outside this range can not be perceived (Rosen and Howell, 2010; Rossing, 2007). Based on this fact, in Zhang et al. (2017a) and Roy et al. (2017), high frequencies are used to generate audio which is inaudible to humans but which is captured and classified by a device. Although these attacks may not fit in our specification of adversarial examples (since humans cannot perceive the generated audio, and therefore cannot judge it as benign either), they introduce the idea of using frequency ranges that are out of the human hearing range in adversarial scenarios.

A different strategy is employed in Qin et al. (2019) and Schönherr et al. (2018), where psychoacoustic models (Zwicker and Fastl, 2013) are used to compute the hearing thresholds of different zones of the clean audio signal, which are used to restrict the perturbation to the least perceptible parts. While this strategy is particularly interesting for individual attacks, since the perturbation can be hidden by taking into account the particularities of a single audio signal, it has several limitations when it comes to universal perturbations, due to the fact that the regions of an audio in which the perception is lower vary drastically depending on the signal.

2.2.3. *Human evaluation*

In Cisse et al. (2017) and Kreuk et al. (2018), an ABX test is performed, which is a standard method to identify detectable differences between two choices of sensory stimuli. In this method a subject is asked to listen to two audios A and B, and afterwards a third audio X, which will be either A or B, randomly selected. The objective of this test is to assess if the user is able to distinguish between A and B. Optimally, the accuracy ratio would be 50%, equal to the probability of selecting randomly between the two choices. In our scenario, the two initial audios A and B would correspond to the clean and perturbed audio (in either order).

In Schönherr et al. (2018), a *Multiple Stimuli with Hidden Reference and Anchor* (MUSHRA) test (Schinkel-Bielefeld et al., 2013) is carried out to perform a subjective assessment of the audio quality of adversarial examples. The goal of the test is to score the quality of perturbed audio signals (*anchors*, e.g., adversarial examples) with respect to the original signal (*hidden reference*, in this context, the original audio).[1] According to the results, the adversarial examples obtained considerably lower scores than the clean audio signals.

In Yakura and Sakuma (2019) and Yuan et al. (2018) the adversarial perturbations are embedded in songs, which can be deployed in the physical world without raising suspicions for human listeners (e.g., in elevators or TV advertisements) to force a target model to understand speech commands. In both works a human evaluation is carried out on Amazon Mechanical Turk to qualitatively assess the detectability of their attacks. According to the results presented by the authors, almost none of the participants perceived speech in the perturbed signals. However, a considerable percentage of people reported that an abnormal noise could be noticed in the songs.

In Qin et al. (2019), where adversarial examples are generated for automatic speech recognition tasks, a human evaluation of their attacks is also carried out on Amazon Mechanical Turk, qualitatively assessing different factors. According to the results reported by the authors, the adversarial examples were judged completely identical to the original samples 76% of the times. However, when enhancing their attacks in order to work in the physical world, the perturbations were considerably more detectable, and the adversarial examples were judged different to the original samples 71% of the times.

Finally, in Alzantot et al. (2018a); Carlini et al. (2016); Du et al. (2020); Gong and Poellabauer (2017); Vaidya et al. (2015), experiments with human subjects are performed with the aim of analyzing their response to the task, in order to assess if the adversarial perturbation has any influence on the responses provided by human listeners. However, no analysis of the perceptual distortion introduced by the perturbations is reported, except in Du et al. (2020), in which the subjects are asked to evaluate the noise level of the audio signals.

It is noteworthy that, although a human evaluation is the most reliable method of studying the extent to which the adversarial perturbations are detectable, it is necessary to appropriately model such criteria using distortion metrics. This would allow us to employ such metrics during the optimization process of the perturbations to minimize the perceptual distortion more effectively, or efficiently comparing results from different attacks in a more standardized way, without the need for human intervention. However, defining such metrics is a challenging task. For these reasons, in comparison to these previous works, rather than using a human evaluation to analyze the audibility of adversarial attacks, we aim to study whether the distortion metrics proposed in the literature agree with the human judgement, in order to assess their suitability for such purposes.

Finally, in all these works, only individual adversarial perturbations have been evaluated. In contrast, in this paper we evaluate universal adversarial perturbations, which require higher amounts of distortion in comparison to individual attacks, and which can not be enhanced for each input individually, making it difficult to mask or hide the perturbations in the input signals. To the best of our knowledge, no prior work has reported a human evaluation of universal audio adversarial perturbations.

2.3. *Summary*

Despite the fact that different distortion metrics have been proposed to measure the distortion levels introduced by audio adversarial perturbations, we found that most of the approaches are not enough to adequately represent the human perception of these attacks, as some of the thresholds or acceptable distortion levels assumed in previous works

[1] It is worth mentioning that the MUSHRA test is mainly used to assess the intermediate quality level of coding systems, whereas for small impairments, which should be the case of audio adversarial perturbations, more suitable tests have been proposed, (ITU, 2015).

(Abdoli et al., 2019; Alzantot et al., 2018a; Carlini and Wagner, 2018; Gong et al., 2019; Gong and Poellabauer, 2017; Neekhara et al., 2019; Yang et al., 2018) do not always guarantee that the perturbations are imperceptible. Therefore, the undetectability of the attacks can be questionable. With this paper, we intend to provide evidence and raise awareness about this. We hope that the results reported may contribute to establish a more thorough measurement of the distortion, and therefore, to a more realistic study of audio adversarial examples.

3. Adversarial examples of speech commands

Our goal is to evaluate the detectability of audio adversarial perturbation, and to determine to what extent the metrics commonly used in the literature agree with the human evaluation. To accomplish this goal, we should first establish a number of stepping stones:

1. Identify a suitable and representative audio task.
2. Identify a model appropriate for the task
3. Collect or identify a dataset to train the model.
4. Using the model, generate the adversarial examples for the task.
5. Estimate the actual fooling rate of the adversarial examples.

3.1. *Selection of the task, model, and dataset*

The task we have selected is *speech command classification* since it is an exemplar machine learning task which is part of the repertoire of extensively used speech-based virtual agents, such as smartphones or smart home assistants. Lightweight speech command recognition models are particularly well suited for tasks requiring a continuous monitoring of audio signals in search of keywords, or for resource-constrained devices, due to the high computational cost required by automatic speech recognition models (Zhang et al., 2017b). In fact, it is common for automatic speech recognition systems to have to be activated by predefined short commands, which are usually recognized by keyword-spotting models (Chen et al., 2014; Sainath and Parada, 2015). Therefore, vulnerabilities in speech command classification models can lead to privacy-issues, security breaches or dangerous malfunctions of voice controlled devices in security critical tasks.

The DNN model we have selected is based on the architecture proposed for small-footprint keyword recognition (Sainath and Parada, 2015). Such architecture has been used in related works on adversarial examples (Alzantot et al., 2018a; Du et al., 2020) and as a baseline model in other research tasks (Warden, 2018; Zhang et al., 2017b). The model takes as input an audio waveform, computes the spectrogram of the signal for different time intervals, and extracts a set of MFCC features for each of them. The resulting two-dimensional representation of the audio signal is fed into a Convolutional Neural Network, composed of two convolutional layers with a ReLU activation function, followed by a fully-connected layer and a softmax layer.

We used the Speech Command Dataset (Warden, 2018), which is a widely used dataset in the study of adversarial attacks for speech recognition systems (Abdoli et al., 2019; Alzantot et al., 2018a; Du et al., 2020; Yang et al., 2018). The dataset is composed of recordings of 30 different spoken commands, provided by a large number of different people. Audio files are stored in a 16-bit WAV file, with a sample-rate of 16kHz and a fixed duration of one second. As in previous publications (Alzantot et al., 2018a; Warden, 2018), we selected the following subset of commands to develop our work: “Yes”, “No”, “Up”, “Down”, “Left”, “Right”, “On”, “Off”, “Stop”, and “Go”. Additionally, We will also consider two special classes: “Unknown” (a spoken command not considered in the previous set), and “Silence” (no speech detected in the audio). The selected setup allows multiple factors to be controlled (e.g., the number of spoken commands is limited yet varied) or fixed (e.g., the length of all the audio samples is equal). This allows us to focus on more relevant factors of speech signals, which highly affect the perceptual distortion of adversarial perturbations, as we show in the paper.

3.2. *Generating* single-class *universal perturbations*

As previously mentioned, we focus on *single-class* universal perturbations (Gupta et al., 2019; Vadillo and Santana, 2019), an attack approach that attempts to generate a single perturbation which is able to fool the model for any input corresponding to a particular class y_i. We decided to focus on universal perturbations because an initial experimentation with individual perturbations (crafted using Deepfool algorithm) led us to the conclusion that the perturbations were undoubtedly imperceptible. This conclusion has been reported before in the literature (Fezza et al., 2019) for the case of image adversarial examples. Therefore, we selected the more challenging task of generating universal perturbations, which requires higher distortion levels. The fact that higher distortion levels are required to generate universal perturbations also increases the necessity to assess their imperceptibility in a rigorous and realistic way, in order to properly study the potential of such attacks. Thus, universal perturbations provide an appropriate setup to evaluate the main objective of our paper: to carry out an analysis of the human perception of audio adversarial perturbations, in order to assess whether the distortion metrics employed in the literature correlate with the human judgment. Moreover, we selected *single-class* universal attacks in order to study in more detail the results on different commands. The particular choice of the class to which the target perturbation is applied is a factor that may influence the perceptual distortion of the perturbations.

The selected attack method is based on the strategy proposed in Moosavi-Dezfooli et al. (2017), a state-of-the-art method to generate universal perturbations based on accumulating individual perturbations created for a set of *training* samples using the Deepfool algorithm (Moosavi-Dezfooli et al., 2016). The Deepfool algorithm is an individual adversarial attack method which, given an input x, aims to find the minimal perturbation capable of changing the classification provided to x by the model, that is:

$$r^* = \min_r ||r||_2 \quad \text{s.t.} \quad f(x+r) \neq f(x). \tag{4}$$

The strategy followed by Deepfool is to employ a linear approximation of the decision boundaries learned by the classifier to efficiently approximate the distance between the input and the closest decision boundary (towards which the input will be pushed). This is done iteratively until the perturbed (i.e., adversarial) input is wrongly classified by the model. Let $f_j(x)$ be the output logit of the model f corresponding to the class y_j when an input x is classified, and x_i the adversarial example at the iteration i ($x_0 = x$). At each iteration, the decision region corresponding to the source class $y_c = f(x)$, which can be formally defined as:

$$R = \bigcap_{j=1}^{k} \{x \in \mathbb{X} : f_c(x) \geq f_j(x)\}, \tag{5}$$

is approximated as:

$$\widetilde{\mathcal{R}}_i = \bigcap_{j=1}^{k} \{x \in \mathbb{X} : f_j(x_i) - f_c(x_i) + \nabla f_j(x_i)^\top x - \nabla f_c(x_i)^\top x \leq 0\}, \tag{6}$$

where $\nabla f_j(x)$ denotes the gradients of the logit corresponding to the class y_j with respect to the input. For the sake of simplicity, let us denote $f'_j = f_j(x_i) - f_c(x_i)$ and $w'_j = \nabla f_j(x_i) - \nabla f_c(x_i)$. Based on the simplified decision boundary model $\widetilde{\mathcal{R}}_i$, the adversarial example is updated using the following rule:

$$x_{i+1} = x_i + \frac{|f'_l|}{||w'_l||_2^2} w'_l, \tag{7}$$

where $l \neq c$ represents the index of the class $y_l \neq y_c$ whose decision boundary is closest to x_i according to the following proximity criterion:

$$l = \text{argmin}_{j \neq c} \frac{|f'_j|}{||w'_j||_2}. \tag{8}$$

In order to generate a universal perturbation v, given a *training set* $\bar{X} = \{\bar{x}_1, \ldots, \bar{x}_n\}$ of n input samples, the UAP algorithm (Moosavi-Dezfooli et al., 2017) iteratively takes an input $\bar{x}_i \in \bar{X}$, computes an individual perturbation r_i for $(\bar{x}_i + v)$ using the Deepfool algorithm (unless $f(\bar{x}_i) \neq f(\bar{x}_i + v)$, that is, unless v is already capable of fooling the model for the input $\bar{x}_i$), and adds the local perturbation r_i to the universal perturbation v:

$$v \leftarrow \mathcal{P}_\epsilon(v + r_i). \tag{9}$$

The projection operator

$$\mathcal{P}_\epsilon(v) = \text{argmin}_{v'} ||v - v'||_2 \text{ subject to } ||v'||_2 \leq \epsilon \tag{10}$$

is used to bound the norm of the universal perturbation after each update. This process is repeated until a stop criterion is met (e.g., a fixed number of steps or passes through the entire *training* set). We use the *UAP-HC* reformulation of this strategy for audio samples, as presented in Vadillo and Santana (2019), where more details about the process to generate the perturbations can be found.

We generated 5 different universal perturbations per class, starting from a different *training* set of 1000 samples in each case. During the crafting process, the universal perturbations were bounded by the ℓ_2 norm, with a threshold value of $\epsilon = 0.1$. In addition, the Deepfool algorithm was limited to a maximum number of 100 iterations. The overshoot parameter of the Deepfool algorithm was set to 0.1. Finally, the *UAP-HC* algorithm was restricted to 5 epochs, that is, 5 complete passes through the entire *training* set.

Table 1 – Effectiveness of the generated single-class universal perturbations.

Class	Max. FR%		Mean FR%	
	Train	Valid	Train	Valid
Silence	23.80	19.46	22.24	19.61
Unknown	72.70	73.06	70.58	73.51
Yes	74.50	74.36	68.26	66.40
No	86.50	83.77	81.48	79.40
Up	84.20	75.45	82.20	74.73
Down	71.50	65.55	68.06	64.51
Left	52.30	49.73	42.20	40.59
Right	68.70	63.82	60.62	56.47
On	76.00	75.65	54.42	53.28
Off	80.10	73.48	75.18	70.85
Stop	61.40	61.82	56.92	57.30
Go	87.80	80.06	86.24	80.90

3.3. *Effectiveness of the perturbations fooling the model*

To measure the effectiveness of the universal perturbations, we compute the percentage of audios for which the prediction changes when the perturbation is applied. We will refer to this metric as fooling ratio (FR) (Moosavi-Dezfooli et al., 2017). The effectiveness of the generated perturbations is shown in Table 1, for the training set (the set of samples used to optimize the universal perturbation) and for the test set (the set of samples used to compute the effectiveness of the attack for inputs not used during the optimization process). Results are shown for the average effectiveness of the 5 perturbations generated for each class, as well as for the one that maximizes the FR on the training set.

According to the results, the generated adversarial examples are highly effective for the majority of the classes, with a maximum FR above 70% for 7 out of 12 classes in both training and test sets. Note that we obtain a considerably high effectiveness also in the class *unknown*, which is composed of a diverse set of spoken commands. However, the hardest class to fool is *silence*, in which the maximum FR is below 25% in both training and test sets. This may be due to the fact that, according to the nature of the audios corresponding to that class, trying to fool the model by adding a small amount of *noise* is a challenging task.

It is important to bear in mind that the effectiveness of a universal perturbation is directly correlated to the distortion amount introduced. We show in Fig. 2, for each class, the way in which the FR increases as the distortion amount introduced by the perturbations increases. These results have been obtained by scaling the magnitude of a universal perturbation v according to two distortion criteria: the ℓ_2 norm of the perturbation and the decibel difference between the perturbation

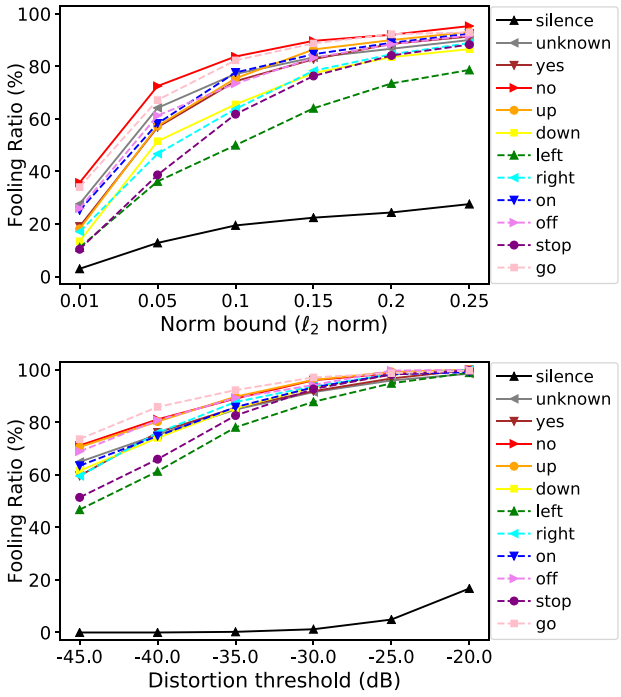

Fig. 2 – Variation in the effectiveness (FR%) in the test set of the generated *single-class* universal adversarial perturbations according to two different criteria: ℓ_2 norm of the perturbation (top) and $dB_{x,max}(v)$ metric with respect to each input signal x (bottom).

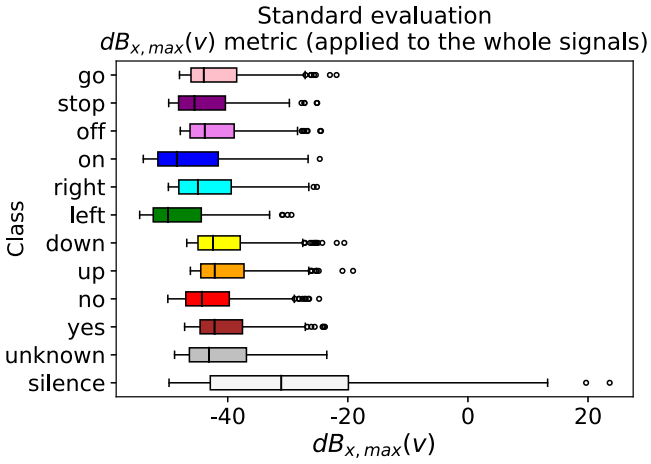

Fig. 3 – Distortion level of the generated *single-class* universal perturbations, evaluated in the test set using the standard evaluation approach: $dB_{x,max}(v)$ applied to the whole signals. Results are averaged for the 5 perturbations generated for each class.

and each sample of the dataset. In the first case, the perturbation signal is scaled in order to ensure that its norm equals the desired threshold, and it is equally applied to every input sample. In the second case, the perturbation signal is scaled for every input sample x, in order to ensure that the $dB_{x,max}(v)$ metric equals the specified threshold.

The fact that the FR is directly correlated with the distortion level implies that there is a trade-off between the effectiveness and the detectability of the attacks. Therefore, to adequately study the risk posed by audio adversarial attacks, it is important to establish realistic and rigorous criteria for assessing the human perception of such attacks.

4. Evaluation of the distortion using similarity metrics

While the ability to fool the model is an essential ingredient of adversarial examples, the other requirement is that the perturbation is not noticed by humans. In this section, we evaluate the distortion produced by the generated adversarial perturbations, according to different criteria.

4.1. *Evaluating the distortion: The standard, uninformed way*

We first computed the distortion according to the standard approaches employed in previous works on adversarial examples in speech recognition tasks (Abdoli et al., 2019; Carlini and Wagner, 2018; Neekhara et al., 2019; Yang et al., 2018), as described in equation (1). Note that according to this metric, the lower the distortion value, the less detectable the perturbation. In Carlini and Wagner (2018), where individual adversarial perturbations are created for speech transcription scenarios, the mean distortion of the generated perturbations is −31dB, and the 95% interval for distortion ranges from −15dB to −45dB.[2] Approximately the same range of distortion is reported in Yang et al. (2018). In Neekhara et al. (2019), where universal adversarial perturbations are generated also for speech transcription models, the distortion level of the perturbations is bounded under different thresholds, obtaining a mean distortion of approximately −42dB in the best case and −30dB in the worst case. In Abdoli et al. (2019), the mean distortion levels of the generated universal perturbations for speech command classification are approximately between −18dB and −25dB. Overall, distortion levels below −32dB are considered acceptable in these works.

Fig. 3 shows the distortion level of the generated perturbation with respect to each input sample in the test set, according to the same approach. Results are computed independently for each class, and averaged for the 5 trials carried out in each of them. Table 2 shows the mean distortion level obtained for each class. As can be seen, the mean distortion is below −40dB in all the classes except *silence*, in which the mean distortion is of −29.52dB.[3] Moreover, without considering the class *silence*, more than 90% of the samples are below −32dB in all the cases. Therefore, our perturbations can be considered as highly acceptable according to this standard.

[2] Different mean distortion values are reported depending on the attack strategy, ranging from -18dB to -45dB.

[3] This effect can be explained by the fact that, due to the nature of the samples corresponding to the class *silence*, their loudness level is lower than for the rest of classes.

Table 2 – Distortion levels produced by the generated single-class universal perturbations (standard evaluation). Results are averaged for the 5 experiments carried out for each class.

Class	Mean $dB_{x,max}(v)$	% of samples below -32dB
Silence	–29.52	48.04
Unknown	–41.35	90.20
Yes	–40.58	90.45
No	–42.56	93.09
Up	–40.24	89.18
Down	–40.63	90.64
Left	–48.10	99.03
Right	–43.30	95.20
On	–46.31	96.21
Off	–42.01	94.03
Stop	–43.92	96.11
Go	–41.88	93.28

4.2. *Evaluating the distortion: Detailed and signal-part-informed way*

In order to measure the distortion in more detail, we employed the approach introduced in Vadillo and Santana (2019). In this case, the distortion induced by the perturbation v in the original sample x is computed in terms of the difference between both the maximum (as defined in Eq. (1)) and the mean decibel values, defined as:

$$dB_{x,mean}(v) = dB_{mean}(v) - dB_{mean}(x), \tag{11}$$

where

$$dB_{mean}(x) = 20 \cdot \log_{10}\left(\frac{1}{d}\sum_{i=1}^{d}|x_i|\right). \tag{12}$$

Furthermore, previous work on evaluating the naturalness of adversarial examples in the audio domain computes the distortion between two signals by applying the metrics to the entire signals (Abdoli et al., 2019; Carlini and Wagner, 2018; Neekhara et al., 2019; Yang et al., 2018). In Vadillo and Santana (2019), the application of both metrics in two different parts of each audio signal is advocated: the *vocal* part and the *background* part. This differentiation is due to the fact that, for spoken commands, the amount of sound outside the vocal part is considerably lower. Thus, the same amount of perturbation would be perceived differently depending on the infected part.

As we are handling short single-command audio signals, the vocal part of an audio signal $x = \{x_1, \ldots, x_d\}$ will be delimited by the contiguous subsequence $\{x_a, \ldots, x_b\}$, $1 \leq a, b \leq d$, containing 95% of the accumulated energy of the signal, that is:

$$\frac{\sum_{i=a}^{b} x_i^2}{\sum_{i=1}^{d} x_i^2} \approx 0.95. \tag{13}$$

Thus, we will assume that the two remaining subsequences $\{x_1, \ldots, x_{a-1}\}$ and $\{x_{b+1}, \ldots, x_d\}$ will be composed just of background noise. Notice that this partition is well suited for single command audios in which it is assumed that the vocal part of the signal is contiguous. Audio signals belonging to the *silence* class will be omitted from the analysis of the vocal part, as they are composed only of background noise, without any vocal part.

The results obtained with the described evaluation approach are shown in Fig. 4. The first row of the figure shows the results obtained using $dB_{x,max}(v)$ metric, and the bottom row the results obtained using $dB_{x,mean}(v)$ metric. Notice the difference between the horizontal axis scales of the figures.

By comparing the perturbations in the vocal part and the background part, it can be seen that perturbations in the vocal part are less noticeable, with a decibel difference significantly lower, which occurs using both $dB_{\max}$ and dB_{mean} distortion metrics.

Regarding the distortion amount in the vocal part, the obtained results are significantly below the threshold of −32dB in almost all the samples, independently of the metric. Compared to the sound intensity level of a normal conversation, a distortion of −30dB corresponds to the weakest audible signal between 10kHz and 100Hz frequency range (Smith, 1997), which is roughly the difference between the ambient noise in a quiet room and a person talking (Carlini and Wagner, 2018).

While the distortion level outside the vocal part is still acceptable under the $dB_{\max}$ metric, according to the dB_{mean} metric the distortion exceeds the threshold of −32dB for a great majority of the samples. In fact, in about half of the cases the difference in decibels is greater than −20dB, which may indicate that the perturbations could be highly detectable in those parts.

5. Human evaluation of voice command adversarial examples

While the methods presented in Section 4.2 provide a more accurate and detailed assessment on the quality of the adversarial examples, the metrics used are not expected to capture all the subtleties of a proper human evaluation. Therefore, we designed an experiment in which human subjects listen to audio adversarial examples and judge them according to different criteria. The main goal of the experiment was to study to which extent the perturbations are detectable by humans. In this section we describe the experimental design and its results.

5.1. *Experimental design*

A set of 36 subjects, independent of the research, was selected to conduct the experiment. Each participant was instructed to listen to different audio clips and answer some questions about them.

The experiment is composed of two parts:

- In the first part, the naturalness of the generated universal adversarial examples is investigated. The other question investigated is to what extent the distortion produced by the perturbation affects the understandability of the spoken commands. To address these questions, each participant is asked to listen to a set of 12 audio clips, six of them

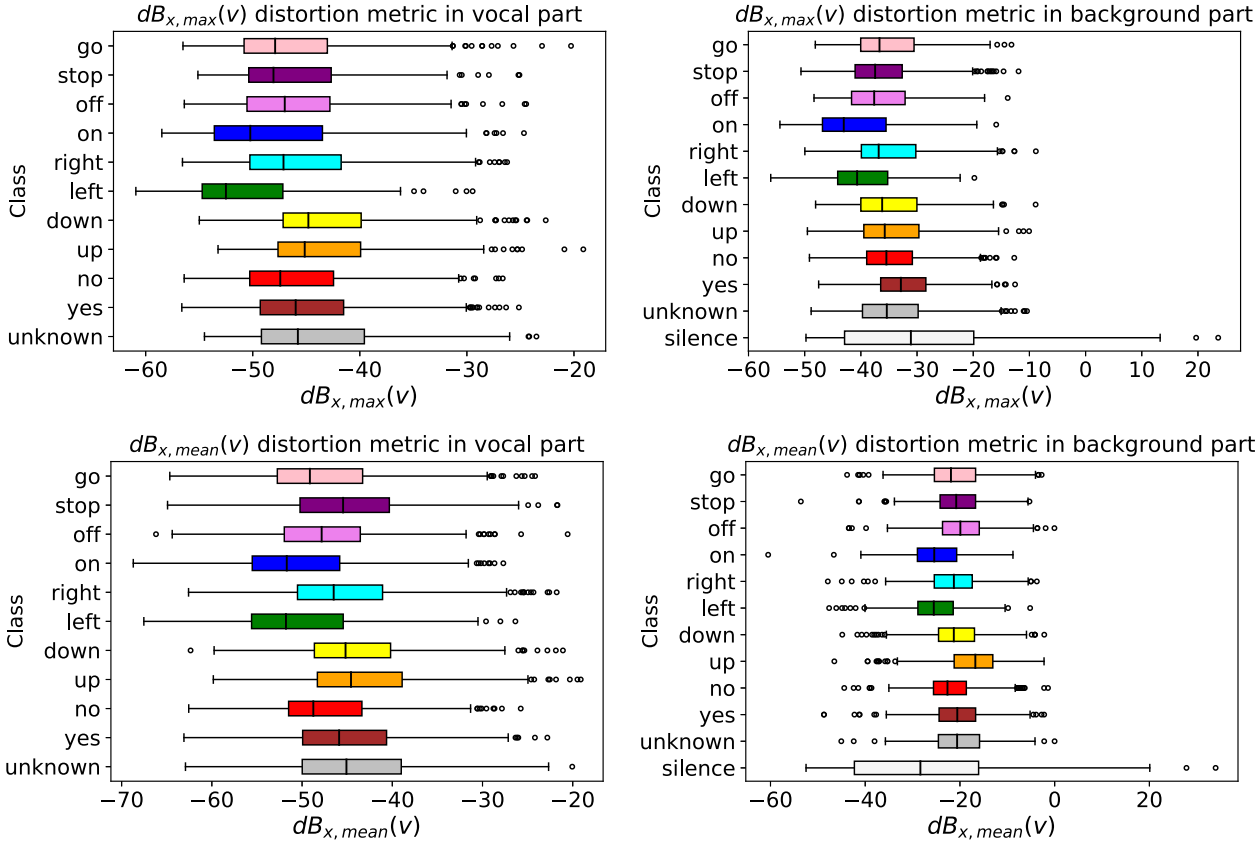

Fig. 4 – Distortion level of the generated *single-class* universal perturbations, evaluated in the test set using $dB_{x,max}(v)$ metric (top row) and $dB_{x,mean}(v)$ metric (bottom row). For each audio, the distortion has been measured in the vocal part as well as in the background part. Results are averaged for the 5 perturbations generated for each class.

clean and the other six adversarially perturbed, and provide the following information[4]:

- Identify the command that can be heard in the audio clip, in order to determine if the adversarial perturbations affect the understandability of the spoken commands.
- Assess the level of naturalness of the audio clip, in order to study whether the adversarial examples are perceived as perturbed audios in comparison to clean instances. As both clean and perturbed audios will be tested, the comparison between the results obtained in both cases may reflect if the perturbations are perceived just as a regular background noise or other ordinary perturbations, or whether they are perceived as artificial or malicious. In the experiment, the subjects evaluated the naturalness on a scale from 1 to 5, with the following scale provided as reference:
  1) Clearly perturbed audio with an artificial sound or noise.
  2) The audio is slightly perturbed by an artificial sound or noise, not likely to be caused by the low quality of the microphones or ambient sounds.
  3) Not sure
  4) No obvious signs of an artificial perturbation. The detectable perturbations are likely to be caused by a low- or mid-quality microphone, ambient sounds or ordinary noises.
  5) The audio clip clearly does not contain any artificial perturbation.
- In the second part of the experiment, each participant performed an ABX test, a method to identify detectable differences between two choices of sensory stimuli. In this method, a subject is asked to listen to two audios A and B, and afterwards a third audio X, which will be either A or B, randomly selected. The goal of the test is to evaluate if the subject is able to determine if X corresponds to A or to B. In our experiment, the two initial audios A and B will correspond to the clean and perturbed audio, in any order. In addition, the participant will be asked to report the confidence of their decision (high, medium and low confidence). Thus, this test will determine if the perturbations are detectable in comparison to the clean audio sample. Six trials were carried out in each experiment, that is, six sets of three audio clips A, B and X.

We included a "catch" ABX trial in the second part of the experiments, in which the three audios were the same, which will be used to discard those experiments in which a high confidence is reported in the classification of X. To prevent the

[4] Participants are instructed to reproduce each audio a maximum of two times.

Table 3 – Summary of the experimental setup designed for the human evaluation of the distortion produced by the universal perturbations.

Experiment	Intensity	Audio samples (part 1)			ABX trials (part 2)
		Clean	Perturbed	Total	
1	Low	6	6	12	6
2	Low	6	6	12	6
3	Low	6	6	12	6
4	Medium	6	6	12	6
5	Medium	6	6	12	6
6	Medium	6	6	12	6
7	High	6	6	12	6
8	High	6	6	12	6
9	High	6	6	12	6

catch ABX trials from biasing the responses of the participants to real trials in any way, we added the catch questions at the end of the tests.

Due to the fact that the audio clips of the dataset contain different characteristics, such as the intensity of the spoken command or the amount of background noise, the perception of a perturbation may change according to these features. For this reason, in order to carry out a more in-depth analysis, we decided to classify the audios considering three levels of intensity: low, medium and high. To the best of our knowledge, no prior work has considered this factor in the assessment of the human perception of audio adversarial perturbations. The dB_{mean} metric presented in Eq. (12) will be used to measure the mean intensity of the original audio signals. According to this metric, 99% of the intensities of the audio samples lie approximately in the decibel range [30,85]. By performing a rough uniform binning of the intensity range (known as equal-width binning in the literature (Dougherty et al., 1995)), the levels were defined as follows:

- Low intensity level: audios with a mean distortion below 50dB.
- Medium intensity level: audios with a mean distortion between 50dB and 70dB.
- High intensity level: audios with a mean distortion above 70dB.

To ensure a uniform representation of the different levels of intensity, each experiment was composed of audio signals of only one of these levels. Nine different experiments were created, (three experiments per intensity level), and each of them was assigned to four different participants, making a total of 36 experiments and participants. A summary of the final experimental setup is provided in Table 3, and the frequency of each command in the first part of the experiment (in which participants are asked to classify the audio samples) is shown in Table 4. We ensured that the model correctly classified the original audio samples but incorrectly classified the adversarial examples.

5.2. *Analysis of the results*

In this section we analyze the results obtained in the experiments. As a summary of the participants, the average age was 24.6 years, with a standard deviation of 6.0 years. Out of the 36 participants, 32 were male (88.9%) and 4 female (11.1%).

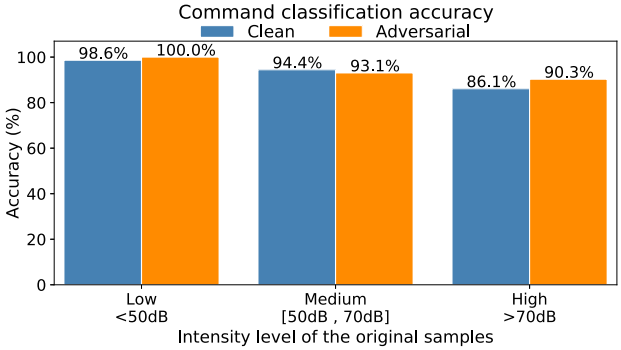

Fig. 5 – Accuracy percentages achieved by the participants of the experiment in the speech command classification task. Results have been split for each sample type (clean or adversarial) as well as for the intensity levels of the original audios in the experiments (low, medium or high).

5.2.1. *Command classification task*

The first factor to be analyzed is the accuracy percentage obtained by humans in the command classification task (first part of the experiment), that is, which percentage of samples have been correctly labeled by humans. The results obtained for each intensity level are summarized in Table 5, which shows the number of wrongly classified audios, and in Fig. 5, which shows the obtained command classification accuracies. In both cases, the results have been computed independently for clean instances and for adversarial examples.

According to the results, the total number of instances wrongly classified considering all the instances, clean and adversarial, is 27 out of 432, which corresponds to a total accuracy in the command classification of $\sim$ 94%. Among all the wrongly classified audios, 15 correspond to the clean samples and 12 to adversarial samples. Overall, these results indicate that the adversarially perturbed spoken commands are as clearly recognizable as the original (clean) audios. In other words, although the adversarial perturbations are able to fool the target model, they do not affect the human understanding of the command. The obtained results are consistent with

Table 4 – Number of audios per command used in the experiments (part 1).

Type	Sil.	Unk.	Yes	No	Up	Down	Left	Right	On	Off	Stop	Go
Low intensity	3	3	3	3	3	3	3	3	3	3	3	3
Medium intensity	3	3	3	3	3	3	3	3	3	3	3	3
High intensity	3	5	3	4	0	4	2	2	3	3	0	7
Clean	7	4	5	5	4	4	4	2	7	4	4	4
Perturbed	2	7	4	5	2	6	4	6	2	5	2	9
Total Frequency	9	11	9	10	6	10	8	8	9	9	6	13

Table 5 – Number of wrongly classified audios (part 1).

Intensity level	Samples		
	Clean	Adv.	All
Low	1	0	1
Medium	4	5	9
High	10	7	17
All	15	12	27

those achieved in Du et al. (2020), where the success rate of a set of people in classifying audio commands is reported using the same dataset as us, but without considering *silence* or *unknown* as classes and without differentiating between the intensity level of the original signals. According to the results reported in Du et al. (2020), the accuracy in recognizing the commands was 93.5% for clean samples and 92.0% for adversarial examples.

Finally, comparing the results for the three intensity levels of the audios, 17 of the wrongly classified audios correspond to the high intensity audios, 9 correspond to the medium intensity and only 1 to the low intensity. These results suggest that, the higher the intensity level of the audio signal, the more difficult it is to correctly identify the spoken commands.

5.2.2. *Naturalness*

The results obtained in the analysis of the naturalness level assigned to the instances is displayed in Fig. 6. The figure shows the frequencies with which samples are classified in each naturalness level, split according to the sample type (clean or adversarial). In addition, the results are jointly computed for all the experiments (top left) as well as for each intensity level individually: low (top right), medium (bottom left) and high (bottom right). Considering all the experiments, it can be observed that the adversarial examples obtained lower scores in comparison to the clean samples. We verified by an exact multinomial statistical test that there exist significant differences regarding the scores assigned to clean and adversarial audios (achieving a p-value below a tolerance of 0.01). Indeed, while 63.0% of the clean samples are classified with a naturalness level of 4 or 5, only 39.8% of adversarial examples have been classified in the same range. These results indicate that, in general, the adversarial perturbations are perceived in the audio signals as artificial sounds or noises with a considerably higher frequency than clean samples.

Doing the same analysis independently for each intensity level, it can be observed that the main difference is given in the lowest intensity level, in which 75.0% of the adversarial examples achieved a score of 1 or 2, while only 4.2% of clean samples were classified in that range. For the highest intensity level, however, the percentage of adversarial examples which scored a 4 or 5 (62.5%) is even greater than the corresponding percentage for clean samples (47.2%). Thus, the human perception of the adversarial examples is clearly related to the intensity level of the original audio signals. This is a remarkable fact that should be taken into consideration in the evaluation of audio adversarial examples.

5.2.3. *ABX test*

In order to better evaluate if the perturbations are perceivable, the results obtained in the ABX test (second part of the experiment) have been analyzed. This is summarized in Fig. 7. The first row of the figure shows the percentage of success cases in the ABX test, that is, the percentage of cases in which the unknown audio (audio X) has been correctly classified. The second row shows the confidence level of the answers. All these results have been computed independently for each intensity level.

The success rate of the experiments with low and medium intensity levels is of 97.2% and 81.9% respectively, revealing that the perturbations are clearly perceivable in such cases. On the contrary, only a 55.6% success rate is achieved for high intensity levels, close to the optimum value of 50%, which is equivalent to a random guessing. We verified by an exact binomial test[5] that the achieved success ratio is not significantly greater (achieving a p-value of $\approx$ 0.2) than the probability $p = 0.5$ corresponding to a binomial distribution $X \sim B(n = 72, p = 0.5)$, where n is the sample size. This fact indicates that, in such cases, the adversarial examples are not distinguishable from their corresponding clean audio samples. It is worth noting that, given our experimental setup, 95% (ClopperPearson) confidence intervals of the success ratio is $[0.90, 0.99]$ for low intensity audios, $[0.71, 0.90]$ for medium intensity audios and $[0.43, 0.67]$ for high intensity audios. The results provided can, therefore, be considered representative of the human perception of the distortion.

Consistently with the success rates, the subjects were highly confident in providing their answers in 90.3% of the cases in the experiments containing audios with low intensity levels, and 62.5% in the experiments containing audios

[5] The alternative hypothesis of the test is that the empirical success ratio is greater than $p = 0.5$. The same test with the alternative hypothesis that the empirical ratio is not equal to 0.5 obtained a p-value of ≈ 0.41.

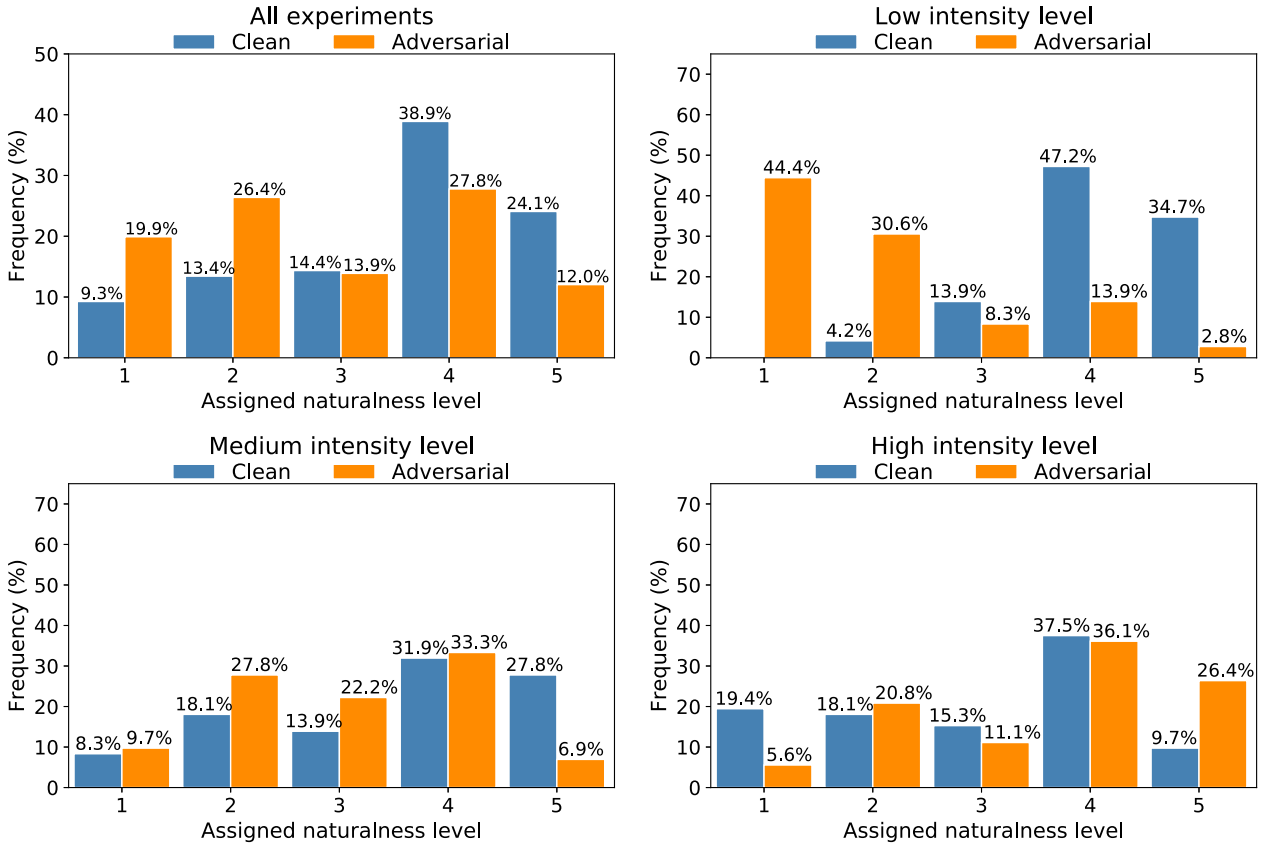

Fig. 6 – Analysis of the naturalness level assigned to the audio samples of the speech command classification task in all the experiments, split by sample type (clean or adversarial). The results are computed for all the experiments (top left) as well as for each intensity level individually: low (top right), medium (bottom left) and high (bottom right).

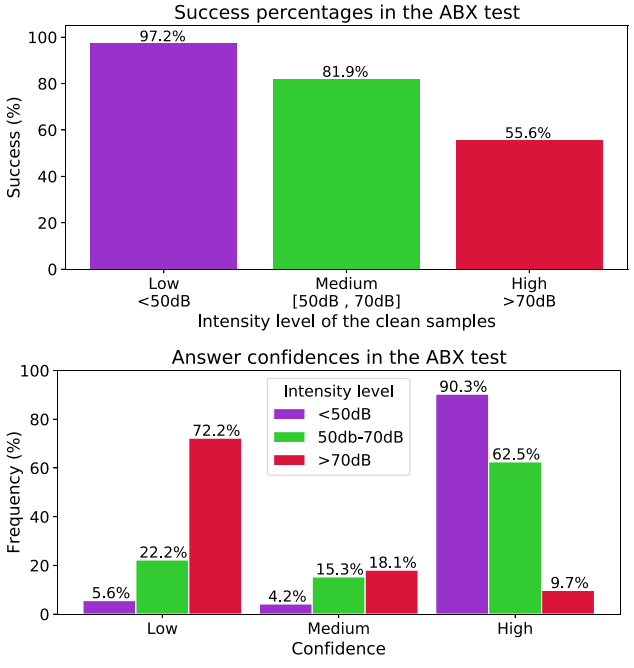

Fig. 7 – Success percentages obtained in the ABX test (top) and confidence levels of the answers in the test (bottom), both computed independently for each intensity level.

with medium intensity levels. Contrarily, only in 9.7% of the answers the participants reported a high confidence in the experiments containing audios with high intensity levels, whereas in 72.2% of the answers a low confidence was reported.

Overall, these analyses demonstrate that the detectability of the perturbations largely depends on the intensity level of the clean audio, being detectable for audios with low and medium intensity levels, but not perceivable for audios with a high intensity level.

It is worth mentioning that, according to the standard approach used in previous related works to measure the detectability of audio adversarial examples, the crafted perturbations were far below the maximum acceptable distortion. However, according to the signal-part-informed approach, particularly evaluating the $dB_{mean,x}(v)$ in the background part of the signal, the distortion exceeded the threshold for a large percentage of samples. Thus, the results obtained in this section reinforce our proposal about the need to employ more rigorous approaches in order to measure and set a threshold on the distortion produced by the adversarial perturbations in a more representative way.

We encourage the reader to listen to some of the adversarial examples generated, to empirically assess the perceptual distortion of adversarial perturbations, as well as to compare the distortion levels obtained using the standard approach

with those obtained using the detailed signal-part-informed approach, to verify that the latter is clearly more correlated with the human judgment.[6]

6. Conclusions

In this paper we have addressed the measurement of the perceptual distortion of audio adversarial examples, which remains a challenging task despite being a fundamental condition for effective adversarial attacks. For this purpose, we have performed an analysis of the human perception of audio adversarial perturbations for speech command classification tasks, and this analysis has been used to assess whether the distortion metrics employed in the literature correlate with the human judgment.

We have found out that, while the distortion levels of our perturbations are acceptable according to the standard evaluation approaches employed by convention, the same perturbations were highly detectable and judged as artificial by human subjects. For this reason, we have proposed the use of a more rigorous framework to measure the distortion in a more comprehensive way, based on a differential analysis in the vocal and background parts of the audio signals, which provide a more realistic evaluation of the perceptual distortion. Our experiments with *single-class* universal perturbations for a set of varied commands also demonstrate that there exist differences regarding the effectiveness of the attacks, related to the relative distortion, and how the perceptual distortion of the perturbations changes depending on the intensity levels of the audio signal in which it is injected.

These results highlight the lack of audio metrics capable of modeling the human perception in a realistic and representative way, and, as a consequence, stress the need to include human evaluation as a necessary step for validating methods used to generate adversarial perturbation in the audio domain. We hope that future works could advance in this direction in order to fairly evaluate the risk that adversarial examples suppose.

As future research, we intend to extend the analysis and methodologies considered in this paper for the evaluation of the perceptual distortion to more complex tasks, such as automatic speech recognition, which can lead to a more comprehensive and uniform framework for the measurement of the perceptual distortion in the audio domain. Although the proposed evaluation framework can be applied to such audio problems, further research may be needed to validate the feasibility of these strategies for these problems, or to develop more particular and suitable frameworks.

In addition, we believe that the findings reported in this paper can be used to generate more imperceptible attacks, for instance, by considering different distortion metrics during the optimization of the adversarial perturbations, with the objective of minimizing the introduced perceptual distortion in the clean samples. Similarly, finding optimization strategies which are capable of integrating complex models of the human hearing system, such as psychoacoustic models, to generate universal perturbations is an interesting research line. Whereas such strategies have been investigated in recent works to generate individual perturbations, it is an open question whether they can be applied to generate more imperceptible universal perturbations, which we hope future work will explore. Nevertheless, to make advances in the field of robust speech recognition, different types of perturbations and attack strategies need to be considered, and it may not be possible, effective or efficient in all cases to make use of such models to generate adversarial perturbations. Therefore, independently of the particular method employed to minimize the distortion during the generation of audio adversarial perturbations, it is necessary to stress the relevance of employing solid approaches to evaluate the distortion introduced to the inputs in order to promote a more rigorous an realistic research in this field.

Finally, although our analysis, and most of the analysis conducted on the use of adversarial perturbations for audio tasks, is based on perturbations that are obtained as an additive transformation of the signal, other types of perturbations (e.g., convolutions) are worth investigating, in search of more effective or more efficient attack approaches.

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.

CRediT authorship contribution statement

Jon Vadillo: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Roberto Santana:** Conceptualization, Methodology, Validation, Resources, Supervision, Writing – original draft.

Acknowledgements

This work was supported by the Basque Government (PRE_2019_1_0128 predoctoral grant, IT1244-19 and project KK-2020/00049 through the ELKARTEK program); the Spanish Ministry of Economy and Competitiveness MINECO (projects TIN2016-78365-R and PID2019-104966GB-I00); and the Spanish Ministry of Science, Innovation and Universities (FPU19/03231 predoctoral grant). The authors would also like to thank to the Intelligent Systems Group (University of the Basque Country UPV/EHU, Spain) for providing the computational resources needed to develop the project, as well as to all the participants that took part in the experiments.

REFERENCES

Abdoli S., Hafemann L.G., Rony J., Ayed I.B., Cardinal P., Koerich A.L.. Universal adversarial audio perturbations. 2019. arXiv preprint:1908.03173.

[6] https://vadel.github.io/adversarialDistortion/AdversarialPerturbations.html.

Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access 2018;6:14410–30. doi:10.1109/ACCESS.2018.2807385.

Alzantot M., Balaji B., Srivastava M.. Did you hear that? adversarial examples against automatic speech recognition. 2018a. arXiv preprint: 1801.00554.

Alzantot M, Sharma Y, Elgohary A, Ho BJ, Srivastava M, Chang KW. Generating natural language adversarial examples. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2018b. p. 2890–6. doi:10.18653/v1/D18-1316.

Athulya M, Sathidevi P, et al. Mitigating effects of noise in forensic speaker recognition. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). IEEE; 2017. p. 1602–6. doi:10.1109/WiSPNET.2017.8300031.

Bayar B, Stamm MC. A generic approach towards image manipulation parameter estimation using convolutional neural networks. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. ACM; 2017. p. 147–57. doi:10.1145/3082031.3083249.

Bayar B, Stamm MC. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security 2018;13:2691–706. doi:10.1109/TIFS.2018.2825953.

Boles A, Rad P. Voice biometrics: Deep learning-based voiceprint authentication system. In: 12th System of Systems Engineering Conference (SoSE). IEEE; 2017. p. 1–6. doi:10.1109/SYSOSE.2017.7994971.

Bosi M, Goldberg RE. Introduction to Digital Audio Coding and Standards, 721. Springer Science & Business Media; 2012. doi:10.1007/978-1-4615-0327-9.

Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner D, Zhou W. Hidden voice commands. In: 25th USENIX Security Symposium (USENIX Security 16); 2016. p. 513–30.

Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW); 2018. p. 1–7. doi:10.1109/SPW.2018.00009.

Chen G, Parada C, Heigold G. Small-footprint keyword spotting using deep neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2014. p. 4087–91. doi:10.1109/ICASSP.2014.6854370.

Cisse M., Adi Y., Neverova N., Keshet J.. Houdini: Fooling deep structured prediction models. 2017. arXiv preprint: 1707.05373.

Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: Machine Learning Proceedings 1995. Elsevier; 1995. p. 194–202. doi:10.1016/B978-1-55860-377-6.50032-3.

Du T, Ji S, Li J, Gu Q, Wang T, Beyah R. SirenAttack: Generating adversarial audio for end-to-end acoustic systems. In: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security; 2020. p. 357–69. doi:10.1145/3320269.3384733.

Dukler Y, Li W, Lin A, Montufar G. Wasserstein of Wasserstein loss for learning generative models. In: Proceedings of the 36th International Conference on Machine Learning. PMLR; 2019. p. 1716–25.

Fang B, Sun F, Liu H, Liu C. 3D human gesture capturing and recognition by the IMMU-based data glove. Neurocomputing 2018;277:198–207. doi:10.1016/j.neucom.2017.02.101.

Feng H, Fawaz K, Shin KG. Continuous authentication for voice assistants. In: Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking. ACM; 2017. p. 343–55. doi:10.1145/3117811.3117823.

Fezza SA, Bakhti Y, Hamidouche W, Déforges O. Perceptual evaluation of adversarial attacks for CNN-based image classification. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE; 2019. p. 1–6. doi:10.1109/QoMEX.2019.8743213.

Gao J, Galley M, Li L, et al. Neural approaches to conversational AI. Foundations and Trends® in Information Retrieval 2019;13:127–298. doi:10.1561/1500000074.

Gong Y, Li B, Poellabauer C, Shi Y. Real-Time Adversarial Attacks. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19); 2019. p. 4672–80. doi:10.24963/ijcai.2019/649.

Gong Y., Poellabauer C.. Crafting adversarial examples for speech paralinguistics applications. 2017. arXiv preprint: 1711.03280.

Gong Y., Poellabauer C.. An overview of vulnerabilities of voice controlled systems. 2018. arXiv preprint: 1803.09156.

Goodfellow I.J., Shlens J., Szegedy C.. Explaining and harnessing adversarial examples. 2014. arXiv preprint: 1412.6572.

Gupta T., Sinha A., Kumari N., Singh M., Krishnamurthy B.. A method for computing class-wise universal adversarial perturbations. 2019. arXiv preprint: 1912.00466.

Hassan MM, Uddin MZ, Mohamed A, Almogren A. A robust human activity recognition system using smartphone sensors and deep learning. Future Generation Computer Systems 2018;81:307–13. doi:10.1016/j.future.2017.11.029.

Heigold G, Moreno I, Bengio S, Shazeer N. End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2016. p. 5115–19. doi:10.1109/ICASSP.2016.7472652.

Huang L, Pun CM. Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced densenet-bilstm network. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2020;28:1813–25. doi:10.1109/TASLP.2020.2998870.

(ITU) I.T.U.. Recommendation ITU-r BS.1116-3, methods for the subjective assessment of small impairments in audio systems. https://www.itu.int/rec/R-REC-BS.1116/en; 2015.

Jordan M., Manoj N., Goel S., Dimakis A.G.. Quantifying perceptual distortion of adversarial examples. 2019. arXiv preprint: 1902.08265

Kereliuk C, Sturm BL, Larsen J. Deep learning and music adversaries. IEEE Transactions on Multimedia 2015;17:2059–71. doi:10.1109/TMM.2015.2478068.

Kreuk F, Adi Y, Cisse M, Keshet J. Fooling end-to-end speaker verification with adversarial examples. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 1962–6. doi:10.1109/ICASSP.2018.8462693.

Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 86–94. doi:10.1109/CVPR.2017.17.

Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2574–82. doi:10.1109/CVPR.2016.282.

Neekhara P, Hussain S, Pandey P, Dubnov S, McAuley J, Koushanfar F. Universal adversarial perturbations for speech recognition systems. In: Proc. Interspeech 2019; 2019. p. 481–5. doi:10.21437/Interspeech.2019-1353.

Nunez JC, Cabido R, Pantrigo JJ, Montemayor AS, Velez JF. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition 2018;76:80–94. doi:10.1016/j.patcog.2017.10.033.

Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, 2015.
Qin Y, Carlini N, Cottrell G, Goodfellow I, Raffel C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proceedings of the 36th International Conference on Machine Learning. PMLR; 2019. p. 5231–40.
Rosen S, Howell P. Signals and Systems for Speech and Hearing, 29. Brill; 2010.
Rossing T. Springer Handbook of Acoustics. Springer Science & Business Media; 2007. doi:10.1007/978-1-4939-0755-7.
Roy N, Hassanieh H, Roy Choudhury R. BackDoor: Making microphones hear inaudible sounds. In: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. ACM; 2017. p. 2–14. doi:10.1145/3081333.3081366.
Sainath TN, Parada C. Convolutional neural networks for small-footprint keyword spotting. In: Interspeech 2015; 2015. p. 1478–82.
Schinkel-Bielefeld N, Lotze N, Nagel F. Audio quality evaluation by experienced and inexperienced listeners. In: Proceedings of Meetings on Acoustics ICA2013. ASA; 2013. p. 060016. doi:10.1121/1.4799190.
Schönherr L., Kohls K., Zeiler S., Holz T., Kolossa D.. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. 2018. arXiv preprint: 1808.05665.
Smith S.W., et al. The scientist and engineer's guide to digital signal processing. 1997.
Snyder D, Garcia-Romero D, Povey D, Khudanpur S. Deep neural network embeddings for text-independent speaker verification. In: Proc. Interspeech 2017; 2017. p. 999–1003. doi:10.21437/Interspeech.2017-620.
Sun Y, Chen Y, Wang X, Tang X. Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems; 2014. p. 1988–96.
Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations; 2014. p. 1–10.
Vadillo J., Santana R.. Universal adversarial examples in speech command classification. 2019. (Submitted for publication). arXiv preprint: 1911.10182.
Vaidya T, Zhang Y, Sherr M, Shields C. Cocaine noodles: Exploiting the gap between human and machine speech recognition. In: 9th USENIX Workshop on Offensive Technologies (WOOT 15); 2015. p. 1–14.
Warden P.. Speech commands: A dataset for limited-vocabulary speech recognition. 2018. arXiv preprint: 1804.03209.
Yakura H, Sakuma J. Robust audio adversarial example for a physical attack. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19); 2019. p. 5334–41. doi:10.24963/ijcai.2019/741.
Yang Z., Li B., Chen P.Y., Song D.. Characterizing audio adversarial examples using temporal dependency. 2018. arXiv preprint: 1809.10875.
Yuan X, Chen Y, Zhao Y, Long Y, Liu X, Chen K, Zhang S, Huang H, Wang X, Gunter CA. Commandersong: A systematic approach for practical adversarial voice recognition. In: 27th USENIX Security Symposium (USENIX Security 18); 2018. p. 49–64.
Zeng J, Zeng J, Qiu X. Deep learning based forensic face verification in videos. In: 2017 International Conference on Progress in Informatics and Computing (PIC). IEEE; 2017. p. 77–80. doi:10.1109/PIC.2017.8359518.
Zhang G, Yan C, Ji X, Zhang T, Zhang T, Xu W. Dolphinattack: Inaudible voice commands. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS). ACM; 2017a. p. 103–17. doi:10.1145/3133956.3134052.
Zhang Y., Suda N., Lai L., Chandra V.. Hello edge: Keyword spotting on microcontrollers. 2017b. arXiv preprint: 1711.07128.
Zwicker E, Fastl H. Psychoacoustics: Facts and models, 22. Springer Science & Business Media; 2013.

Jon Vadillo received a B.S. degree in Computer Science from the University of the Basque Country UPV/EHU, Spain, in 2018. He received an M.Sc. degree at the same university, in 2019. He is currently pursuing a Ph.D. degree with the University of the Basque Country UPV/EHU, Spain, in the field of Adversarial Machine Learning. His current research interests comprise the vulnerabilities of Deep Learning models in adversarial scenarios and the application of such technologies in the audio domain.

Roberto Santana received an M.Sc. degree in Computer Science from the University of Havana, Cuba, in 1996. He received a Ph.D. in Mathematics from the University of Havana in 2005 and a Ph.D. in Computer Science from the University of the Basque Country UPV/EHU, in Spain, in 2006. He is a Tenured researcher at the Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Spain. His research interests comprise Machine Learning methods, Evolutionary Algorithms, and Neuroinformatics.