Towards Reliable Data-Driven Sound Recognition Models: Developing Attack and Defense Algorithms

by

Mohammad ESMAEILPOUR

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 20, 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Patrick Cardinal, Thesis Supervisor
Department of Computer Engineering and Information Technology, École de Technologie Supérieure

Mr. Alessandro Lameiras Koerich, Co-supervisor
Department of Computer Engineering and Information Technology, École de Technologie Supérieure

Mr. Éric Granger, President of the Board of Examiners
Department of Systems Engineering, École de Technologie Supérieure

Mr. Christian Desrosiers, Member of the Jury
Department of Systems Engineering, École de Technologie Supérieure

Mr. Douglas O'Shaughnessy, External Examiner
Institut National de Recherche Scientifique, Centre Énergie, Matériaux, Télécommunications

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON DECEMBER 07, 2021

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to sincerely thank both my supervisors, namely Prof. Patrick Cardinal and Prof. Alessandro Lameiras Koerich for their patience, kindness, support, and guidance which added remarkably to my Ph.D. experience. They have been always there for me and without them, it would never be possible to pursue my educational and research careers.

Heartfelt thanks to all my fellow lab-mates at LIVIA for creating an intellectual environment and making good memories during the last four years. I would also thank both undergraduate and graduate students who helped me to practice my mentorship skills.

I wish to express my deepest gratitude to all my family members, especially my parents for their unconditional overwhelming love and endless supports. They have always paved my way to run after my dreams. I can never thank them enough for making me the way I feel today.

Many thanks to my thesis committee and jury members, particularly Prof. O'Shaughnessy and Prof. Desrosiers for reviewing this thesis.

Last but not the least, I would like to acknowledge the valuable support of the Natural Sciences and Engineering Research Council of Canada (NSERC), MITACS, IVADO, and Desjardins under grant agreement number RGPIN 2016- 04855, RGPIN 2016-06628, and IT25105.

Vers des modèles fiables de reconnaissance de sons basés sur des données: développement d'algorithmes d'attaque et de défense

Mohammad ESMAEILPOUR

RÉSUMÉ

La classification des sons environnementaux (CSE) et la reconnaissance automatique de la parole (RAP) ont toujours suscité un intérêt croissant de la part de l'industrie et du monde universitaire en raison de leur vaste gamme d'applications pratiques dans la vie réelle. Par exemple, les réseaux de capteurs multimédias, les systèmes de surveillance et les applications d'assistance vocale intégrées dans nos smartphones utilisent principalement les modèles CSE et RAP. Compte tenu des progrès significatifs réalisés au cours des dernières décennies, la précision de reconnaissance des classificateurs de pointe introduits dans ces domaines a atteint de manière compétitive le niveau de compréhension humain. Cependant, ces modèles de l'état de l'art basés sur des données sont extrêmement vulnérables aux signaux adverses qui sont soigneusement conçus pour tromper les classificateurs vers des sorties incorrectes. Techniquement, un signal contradictoire comporte une légère perturbation qui peut être obtenue par une formulation d'optimisation, et il force le modèle de reconnaissance à prédire des sorties incorrectes prédéfinies par un adversaire. Cela pose un problème de sécurité majeur puisque les signaux adverses ne sont pas non plus détectables par des évaluations subjectives. De plus, ces signaux malveillants sont transférables de manière bijective à des représentations 1D (c'est-à-dire le coefficient cepstral de fréquence Mel - MFCC) et 2D (spectrogrammes 2D) telles que les transformées de Fourier à court terme et les transformées en ondelettes discrètes. Étant donné que la majorité des modèles avancés de CSE et de RAP sont formés sur des représentations, de tels spectrogrammes adverses peuvent effectivement diminuer la précision de reconnaissance de ces modèles. Malheureusement, il existe très peu de recherches sur la défense des classificateurs contre une variété d'attaques adverses ciblées et non ciblées. De plus, ces approches ne sont peut-être pas assez fiables pour protéger les modèles contre les attaques de type boîte blanche et boîte noire.

Comme il n'existe pas de définition standard de la fiabilité d'un algorithme de défense adversariale, nous définissons nos propres implications de la fiabilité et imposons trois conditions principales. Premièrement, un algorithme de défense fiable doit éviter toute opération de filtrage susceptible d'obscurcir les informations du gradient ou de briser la matrice jacobienne. Deuxièmement, il devrait faire un compromis raisonnable entre la précision de reconnaissance, la robustesse contre les attaques adverses (taux de tromperie) et la complexité de calcul de l'algorithme afin de fonctionner en temps réel. Troisièmement, il doit être conçu pour produire un classificateur intrinsèquement fort afin de maximiser le coût de l'attaque (par exemple, le nombre total de calculs de gradient requis) pour l'adversaire. De plus, le respect de chacune de ces conditions ne doit pas entrer en conflit avec une autre. Dans cette thèse, nous développons des algorithmes de défense et d'attaque fiables pour les systèmes CSE et RAP avancés de bout en bout et au niveau des représentations, organisés en quatre chapitres et cinq annexes.

Notre première contribution est le développement d'un classificateur CSE principalement en ce qui concerne les conditions de fiabilité de notre troisième défense. Plus précisément, nous concevons un classificateur basé sur un ensemble dans la partie frontale, car il est plus robuste contre les attaques adverses. En outre, nous exploitons un réseau antagoniste génératif (RAG) avec des architectures optimisées pour les réseaux générateur et discriminateur en arrière-plan pour l'augmentation du spectrogramme. Nous démontrons que ce cadre de classification surpasse d'autres architectures conventionnelles (par exemple, les machines à vecteurs de support) et celles basées sur l'apprentissage profond sur des ensembles de données CSE de référence.

En deuxième lieu, nous développons une approche robuste pour sécuriser les modèles CSE contre un large éventail d'attaques adverses de type boîte blanche et boîte noire. Cet algorithme respecte toutes les conditions de fiabilité de la défense susmentionnées et il réalise un compromis très raisonnable entre la précision de reconnaissance et le taux de tromperie des attaques. En outre, nous étudions le rapport de transférabilité des attaques adverses entre les classificateurs classiques et ceux basés sur les réseaux neuronaux. En fonction de ces résultats, nous avons reconfiguré notre configuration dorsale pour combler l'écart entre la robustesse contre les attaques et les performances du classificateur frontal. Par exemple, nous avons utilisé un filtrage Highboost, une opération de réduction de la dimensionnalité, diverses visualisations de spectrogrammes logarithmiques et un codeur automatique de débruitage convolutif. Les expériences que nous avons menées sur quatre ensembles de données difficiles corroborent les performances supérieures de notre approche de défense par rapport aux autres algorithmes.

Notre troisième contribution est la caractérisation expérimentale de la relation inverse entre la précision de reconnaissance et la robustesse du classificateur de la victime contre les attaques adverses ciblées et non ciblées. De plus, nous identifions quelques paramètres du spectrogramme qui contribuent à maximiser le coût de l'attaque pour l'adversaire. Ceci est tout à fait conforme à notre troisième condition de fiabilité qui oblige à développer un classificateur de reconnaissance intrinsèquement fort. Ces paramètres doivent être appliqués avant la production du spectrogramme, de sorte qu'ils n'affectent pas négativement la distribution de la matrice jacobienne, que ce soit pendant l'entraînement ou l'exécution.

Comme quatrième contribution, nous développons une approche de défense haut de gamme pour les systèmes RAP de bout en bout, en particulier les modèles de transcription de la parole-en-texte. Cet algorithme est basé sur la synthèse d'un nouveau signal en utilisant la distance d'accord ajustée et il répond entièrement à nos conditions de fiabilité de défense prédéfinies. Nous utilisons un RAG multi-discriminateur avec de nouvelles architectures convolutionnelles avec résidu pour les réseaux générateur et discriminateur. Ensuite, nous entraînons ce modèle génératif dans l'espace de Sobolev car il est étroitement lié aux séries de coefficients de Fourier comme le MFCC. En outre, nous proposons une nouvelle technique de contrainte pour le réseau générateur afin d'améliorer sa stabilité et sa généralisation pendant l'entraînement et l'exécution en temps réel, respectivement. Nous avons mené nos expériences contre des attaques adverses de type boîte blanche et boîte noire qui ont été évaluées sur les systèmes de transcription avancés DeepSpeech, Kaldi et Lingvo. Ces expériences indiquent que l'algorithme de défense proposé

surpasse les autres approches en termes de taux d'erreur sur les mots et de précision au niveau des phrases.

Le reste de nos contributions qui ont été publiées dans les conférences et letters du journals phares du traitement du signal sont organisées en annexes. Elles comprennent quatre algorithmes de défense et un algorithme d'attaque contradictoire développés pour les systèmes CSE et RAP. Notre principale motivation pour le développement d'un algorithme d'attaque adversarielle est l'introduction d'une attaque rapide et robuste à exploiter dans le cadre de la défense, comme l'entraînement contradictoire.

Mots-clés: classification des sons environnementaux, reconnaissance de la parole, transcription de la parole-en-texte, attaque adversariale, défense adversariale, spectrogrammes, décomposition de Schur généralisée, distance d'accord, sous-espace adversatif, réseau antagoniste génératif.

Towards Reliable Data-Driven Sound Recognition Models: Developing Attack and Defense Algorithms

Mohammad ESMAEILPOUR

ABSTRACT

Environmental sound classification (ESC) and automatic speech recognition (ASR) have always attracted increasing interest from industry and academia due to their extensive range of practical applications in real-life. For instance, multimedia sensor networks, surveillance systems, and voice assistance applications embedded into our smartphones unanimously employ ESC and ASR models. Regarding the significant progress made over the last few decades, the recognition accuracy of the cutting-edge classifiers introduced in these domains has competitively reached to human-level of understanding. However, these state-of-the-art data-driven models are intensely vulnerable against adversarial signals, which are carefully crafted to fool the classifiers toward any incorrect output phrases. Technically, an adversarial signal carries a slight perturbation achievable through an optimization formulation, and it forces the recognition model to predict incorrect outputs as predefined by an adversary. This poses a major security concern since adversarial signals are not detectable by subjective evaluations either. Moreover, these malicious signals are bijectively transferable to both 1D (i.e., Mel-frequency cepstral coefficient - MFCC) and 2D representations (2D spectrograms) such as short-time Fourier and discrete wavelet transforms. Since the majority of the advanced ESC and ASR models are trained on representations, hence such adversarial spectrograms can effectively debase the recognition accuracy of these models. Unfortunately, there is a limited number of investigations on defending classifiers against various targeted and non-targeted adversarial attacks. Additionally, these approaches might not be reliable enough to secure models against strong white and black-box attacks.

Since there is no standard definition for the reliability of an adversarial defense algorithm, we define our implications from reliability and impose three main conditions. Firstly, a reliable defense algorithm should avoid any filtration operations resulting in obfuscating gradient information or shattering the Jacobian matrix. Secondly, it should make a reasonable trade-off among recognition accuracy, robustness against adversarial attack (fooling rate), and the algorithm's computational complexity to work in real-time. Thirdly, it should be designed to yield an inherently strong classifier to maximize the cost of attack (e.g., the total number of required gradient computations) for the adversary. Moreover, complying with each of these conditions should not conflict with another. This thesis develops reliable defense and attack algorithms for the advanced end-to-end and representation-level ESC and ASR systems organized into four chapters and five appendices.

Our first contribution is developing an ESC classifier mainly in regard to our third defense reliability conditions. More specifically, we design an ensemble-based classifier in the front-end since it is more robust against adversarial attacks. Furthermore, we exploit a generative adversarial network (GAN) with optimized architectures for both the generator and discriminator networks in the back-end for spectrogram augmentation purposes. We demonstrate that this

classification framework outperforms other conventional (e.g., support vector machines) and deep learning-based architectures on benchmarking ESC datasets.

As a second contribution, we develop a robust approach for securing ESC models from a wide range of white and black-box adversarial attacks. This algorithm complies with all the defense reliability conditions mentioned above, and it makes a reasonable trade-off between recognition accuracy and attack fooling rate. Moreover, we study the adversarial transferability ratio between conventional and neural network-based classifiers. According to these findings, we reconfigured our back-end configuration to fill the gap between robustness against attacks and the performance of the front-end classifier. For instance, we employed highboost filtering, dimensionality reduction operation, various logarithmic spectrogram visualizations, and convolutional denoising autoencoder. Our conducted experiments on four challenging datasets corroborate the superior performance of our defense approach compared to other algorithms.

Our third contribution is experimentally characterizing the inverse relation between the recognition accuracy and robustness of the victim classifier against targeted and non-targeted adversarial attacks. Additionally, we identified a few spectrogram settings that maximize the adversary's cost of attack. This is completely in line with our third reliability condition which obliges us to develop an inherently strong recognition classifier. These settings should be applied before spectrogram production; therefore they do not negatively affect the Jacobian matrix's distribution either during training or runtime.

As a fourth contribution, we develop an upscale defense approach for end-to-end ASR systems, particularly speech-to-text transcription models. This algorithm is based on synthesizing a new signal using the adjusted chordal distance, and it entirely meets our predefined defense reliability conditions. We employ a multi-discriminator GAN with novel residual-convolutional architectures for the generator and discriminator networks. Then, we train this generative model in the Sobolev space since it is closely related to coefficients of Fourier series, such as Mel-frequency cepstral coefficients (MFCC). Furthermore, we propose a new constraining technique for the generator network to improve its stability and generalizability during training and real-time execution, respectively. Finally, we run our experiments against white and black-box adversarial attacks benchmarked on the advanced DeepSpeech, Kaldi, and Lingvo transcription systems. These experiments indicate that our proposed defense algorithm outperforms other approaches both in terms of word error rate and sentence-level accuracy.

The rest of our contributions published in the flagship signal processing conference and journal letters are organized into appendices. They include four defense and one adversarial attack algorithm developed for both ESC and ASR systems. Our main motivation for developing an adversarial attack algorithm is introducing a fast and robust attack for exploiting in the reliable defense frameworks such as adversarially training.

Keywords: environmental sound classification, speech recognition, speech-to-text transcription, adversarial attack, adversarial defense, spectrograms, generalized Schur decomposition, chordal distance, adversarial subspace, generative adversarial network.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHMS

Page

Algorithms mentioned in the appendices, namely Algorithm I-1 and V-1, are not indexed herein.

LIST OF ABBREVIATIONS

A-GAN	Autoencoder generative adversarial network
ASR	Automatic speech recognition
BBG	Black-blue-green representation
BIM	Basic Iterative Method
BU	Bayesian uncertainty
CC-DGAN	Class-conditional defense generative adversarial network
CDA	Convolutional denoising autoencoder
CD-GAN	cyclic defense generative adversarial network
CIMP	Cramér integral probability metric
CIR	Channel impulse response
CNN	Convolutional neural network, a.k.a. ConvNet
CRP	Cross recurrence plot
CTC	Connectionist temporal classification
CWA	Carlini and Wagner attack, a.k.a. C&W
DA	Data augmentation
DCT	Discrete cosine transform
DNN	Deep neural network
DWT	Discrete wavelet transform
EA	Evasion attack

EOT	Expectation over transformation
ESC	Environmental sound classification
ENH	Enhanced spectrogram
FGSM	Fast gradient sign method
GAA	Genetic algorithm attack
GAN	Generative adversarial network
GMM	Gaussian mixture model
HMM	Hidden Markov model
IPM	Integral probability metric
JSMA	Jacobian-based Saliency Map Attack
KD	Kernel density
LBFGS	Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Szegedy
LFA	Label Flipping attack
LID	Local intrinsic dimensionality
LIVIA	Le Laboratoire d'imagerie, de vision et d'intelligence artificielle
LLR	Log-likelihood ratio
LSTM	Long short-term memory
MFCC	Mel-frequency cepstral coefficient
PESQ	Perceptual evaluation of speech quality
PG	Purple-gold representation

RBF	Radial basis function
RF	Random forest
RIR	Room impulse response
RNN	Recurrent neural network
SIFT	Scale-invariant feature transform
SKM	Spherical K-means
SLA	Sentence-level accuracy
SM	Surrogate model
STFT	Short-time Fourier transform
STOI	Short-term objective intelligibility
SURF	Speeded-up robust feature
SVD	Singular value decomposition
SVM	Support vector machines
SNR	Signal to noise ratio
QZ	The generalized Schur decomposition, a.k.a. QZ decomposition
WCCGAN	Weighted cycle-consistent Generative adversarial network
WER	Word error rate
WB	White-black representation

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$a(t)$	Continuous signal
α	Scalar value
b	Bias term
ϵ	Optimization threshold
ζ	Audible threshold
$r_i(\cdot)$	Distance between a spectrogram and its nearest neighbors
δ	Adversarial perturbation
ω	Frequency component
s	Scale of transformation
τ	Room filter set
$G(\cdot)$	Generator network
$\mathbf{z}_i$	Random input vector for the generator
$\mathbf{z}_i^*$	Optimized safe vector for the generator
$D(\cdot)$	Discriminator network
$G_{S \to T}$	Generator for mapping source to target
$G_{T \to S}$	Generator for mapping target to source
$\mathcal{L}_{GAN}$	Loss function for the generative adversarial network
$\mathcal{L}_{total}$	Total loss function in a cyclic setup
$\mathcal{L}_{ctc}$	Connectionist temporal classification loss

$l(\cdot)$	Loudness metric, a.k.a. distortion condition
$\ell_{net}(\cdot)$	Cross entropy loss
ℓ_m	Loss function for masking threshold
$\mathbb{E}(\cdot)$	Statistical expectation
$p_r(\cdot)$	Probability distribution of real samples
$p_g(\cdot)$	Probability distribution of the generator network
$\det(\cdot)$	Determinant operation
$H(\cdot)$	Hann window function
$\vec{x}_{\text{org}}$	Original signal
$\vec{x}_{\text{adv}}$	Adversarial signal
$\vec{x}_c$	Candidate adversarial signal
$\mathbf{x}_{\text{org}}$	Original spectrogram
$\mathbf{x}_{\text{adv}}$	Adversarial spectrogram
$\mathbf{y}$	Ground-truth phrase
$\hat{\mathbf{y}}$	Incorrect target phrase
π_i	String tokens without duplication
$\text{sign}(\cdot)$	Sign function
$S_{map}(\cdot)$	Saliency map function
$CF(\cdot)$	Cepstral filtering
$f^*(\cdot)$	Post-activation function

$\mathcal{F}$	Function class
$f(\cdot)$	Critic function
$\mu(\cdot)$	Dominant probability density function
$J(\cdot)$	Jacobian matrix
w	Weight vectors
W^{k_s,p_s}	The Sobolev space with parameters k_s and p_s
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_{\text{Lip}}$	Lipschitz continuity
$\|f\|_{\text{Hil}}$	Hilbert continuity
$\Phi(\cdot)$	Kernel function
M	Pairs of microphone-speaker alignment
h	Room impulse response filter
H_{dim}	Set of room impulse response filters with dimension dim
$\text{chord}(\cdot)$	Chordal distance
λ	Eigenvalue vector
$\text{span}(\cdot)$	Span of the matrix
$\mathcal{X}$	Compact open subset in $\mathbb{R}^{d_z}$
Γ	Random unit-ball
$\Phi^{\mathcal{H}}$	Conjugate transpose for the unit-ball of Φ
$\gamma(\cdot)$	Adjustment for the chordal distance

$\gamma^*(\cdot)$	Optimized adjustment for the chordal distance
$\mathrm{rank}(\cdot)$	Rank of the matrix
R_{ϖ}	Orthogonal regularization
$\Omega_s(\cdot)$	Restricted Sobolev space

INTRODUCTION

Environmental sound classification (ESC) and automatic speech recognition (ASR, i.e., speech-to-text transcription) have always been closely related and active research areas among the signal processing communities. Technically, they have been in development for over half a century and this is presumably due to their vast applications in real-life. For instance, towards analyzing the surrounding scene either for surveillance (Valenzise, Gerosa, Tagliasacchi, Antonacci & Sarti, 2007; Radhakrishnan, Divakaran & Smaragdis, 2005; Cristani, Bicego & Murino, 2004) or multimedia sensor networks (Steele, Krijnders & Guastavino, 2013), there is a constant need to recognize environmental sounds. Moreover, ESC models play an important role in smart acoustic sensor network development (Mydlarz, Salamon & Bello, 2017), IoT-based services (Shah, Tariq & Lee, 2019), smart city safety (Ciaburro & Iannace, 2020; Shah, Tariq & Lee, 2018), and context-aware computing (Chandrakala, Venkatraman, Shreyas & Jayalakshmi, 2021; Chu, Narayanan & Kuo, 2009a; Toffa & Mignotte, 2020).

Likewise, there are numerous applications for ASR systems. In particular, nowadays almost all the smartphones are equipped with standard voice command applications (e.g., Siri, Cortana, Bixby, etc.), which rely on at least a built-in ASR model. Moreover, these models have been recently embedded into devices, which should work under considerable amount of surrounding noises in adverse scenarios (in the presence of environmental sounds). Such devices include but are not limited to home appliances (e.g., smart TVs) and driver's voice assist system in vehicles, which are often involved in noisy environments. ASR systems should efficiently process human or machine-produced speech signals and are designed to enhance user experience. Therefore, the performance of these recognition models significantly matters since they should work in real-time under any scenarios.

Over the past decades, huge improvements have been achieved for both ESC and ASR, especially after the proliferation of deep learning algorithms. Taking only the last decade into account, we

notice a large volume of publications on developing data-driven classifiers, which have gradually reached to the current state of competitiveness to the human level of understanding (Boddapati, Petef, Rasmusson & Lundberg, 2017; Mozilla-DeepSpeech, 2017).

According to the literature, there are two categories of papers in developing data-driven models for ESC and ASR:

1. end-to-end;
2. representation-level (a.k.a. spectrogram and frequency-level feature vectors).

In the first category, often raw 1D audio signals are used for training classification models (Thomae & Dominik, 2016; Tokozume & Harada, 2017; Huang & Leanos, 2018). This usually imposes computational overhead on the learning algorithms since 1D signals have high dimensionality. On the other hand, algorithms that fit in the second category utilize one of the following standard representations[1] for training, namely Mel-Frequency cepstral coefficient (MFCC) (Dave, 2013), short-time Fourier transform (STFT) (Benesty, Chen & Habets, 2011), and discrete wavelet transform (DWT) (Tan, Lang, Schroder, Spray & Dermody, 1994). Since the two latter representations yield the power spectrum of the given input signal, they are often called 2D spectrograms. In the big picture, most of the published works on both ESC and ASR fit in the spectrogram-level category. This is because the devised algorithms require much fewer training parameters than the end-to-end counterpart. Furthermore, the highest recognition accuracy have been often reported for algorithms trained on spectrograms or MFCCs (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Qian, Schwarz et al., 2011; Mozilla-DeepSpeech, 2017; Piczak, 2015a; Shen, Nguyen, Wu, Chen, Chen, Jia, Kannan, Sainath, Cao, Chiu et al., 2019).

[1] We are aware that there are many standard representations for audio and speech signals. However, according to the literature, MFCC, STFT, and DWT are the most popular representations.

During the last few years, the major focus has been designing new architectures such as variants of convolution (Sainath, Mohamed, Kingsbury & Ramabhadran, 2013; Chan, Park, Lee, Zhang, Le & Norouzi, 2021), attention (Bahdanau, Chorowski, Serdyuk, Brakel & Bengio, 2016; Luo, Zhang, Lei & Xie, 2021), and recurrent configurations (Graves, Mohamed & Hinton, 2013; Lee, Kang, Cheon, Kim & Kim, 2021) to improve accuracy and generalizability of the ESC and ASR models. However, it has been demonstrated that these advanced algorithms might undergo extreme vulnerability against carefully crafted adversarial signals both in 1D and 2D spectrogram domains (Carlini & Wagner, 2018; Huang, Lin, Lee & Lee, 2021).

In terms of distribution and acoustic characteristics, an adversarial signal is very similar to the original samples. However it is optimized to fool the model to predict incorrect phrase(s). Unfortunately, those crafted signals are capable enough to debase the performance of all the data-driven models from conventional, namely Kaldi[2] (Povey *et al.*, 2011), to modern such as DeepSpeech[3] (Mozilla-DeepSpeech, 2017) and advanced ESC classifiers. Motivated by this concern, we decided to conduct our research towards addressing the threat of adversarial attacks.

Problem Statement

Nowadays, many companies provide online services to their customers through automated assistant machines, which are able to process human languages and hold conversation fluently[4]. The security protocol of these machines is based on recognizing customer's voice which is known as the Voice-ID technology (Keane, 2010; Kaur, Sandhu, Gera, Kaur & Gera, 2020). For commands such as applying for a new credit card, deactivating a debit card, and ordering a mobile SIM-card, customers can directly communicate with automated machines. This saves time, energy, and human resource. On the other hand, this technology has led to varieties

[2] Kaldi uses hidden Markov models (HMMs).

[3] DeepSpeech uses a sequence of long short-term memory (LSTM) units.

[4] Such as Desjardins bank, Fido (Rogers) communication corporation in Canada and AT&T telecommunication company in USA.

of motivations for hackers to take advantage of this online service. In fact, they can attack the Voice-ID systems (using adversarial signals) to run after stealing private and personal information of the customers and pursuing extortion. Unfortunately, this can be very harmful both for companies and their customers.

The major focus of this thesis is developing approaches for defending state-of-the-art ESC and ASR models against varieties of white and black-box[5], as well as targeted and non-targeted[6] adversarial attacks. This is a key step concerning the development of accurate data-driven signal classifiers for many relevant applications. Our ultimate goal in the current work is effectively moving toward the development of reliable defense algorithms. Unfortunately, to the best of our knowledge, there is neither consensus nor standard definition for defense's reliability. However, our implication from reliability includes at least one of the following conditions[7]:

1. *Avoiding Gradient Obfuscation:* obfuscating gradients (intentionally blocking the normal flow of gradients during training Athalye, Carlini & Wagner (2018b)) has been a common issue with almost all the introduced adversarial defense algorithms thus far. Athalye *et al.* (2018b) have initially characterized this issue which is also known as shattering the Jacobian matrix distribution. They have demonstrated that defense algorithms based on filtering[8] the input samples aiming at removing the potential adversarial perturbation, unanimously provide false senses of security against all types of attacks. Moreover, a simple approximation of the model's post-activation function (Tan & Motani, 2020)[9] reduces the performance of such defense algorithms to almost zero. Thus, one of our main conditions

[5] In a white-box attack, the adversary has access to the victim model, architecture, dataset, training hyperparameters, etc., while there are no such accesses in the black-box scenario.

[6] When the attack algorithm is optimized toward a specific incorrect output, it is called targeted. Attack algorithms that implement optimization formulations toward any incorrect outputs other than the ground-truth are considered non-targeted.

[7] Note that complying with each of these conditions should not violate another.

[8] Direct pre/post-processing operations on the input samples.

[9] This is known as backward-pass differentiable approximation method (Athalye *et al.*, 2018b).

in developing an adversarial defense approach is avoiding any variation of direct filtration so that at least to some extent, circumvent gradient obfuscation. We believe that complying with this condition yields to a correct sense of security.

2. *Making Reasonable Trade-offs Among Model Accuracy, Attack Success Rate, and Computational Cost:* according to Athalye *et al.* (2018b), there are two groups of reliable defense approaches that also relatively (partially) meet the first condition:
   a. *adversarially training-based:* these algorithms augment the original training dataset with adversarial signals in order to force the classifier to learn distributions of the crafted signals (Goodfellow *et al.*, 2015). This is a solid defense strategy, however it has two destructive side-effects:
      – it negatively affects the recognition accuracy of the model,
      – it imposes too much computational overhead on the training procedure since crafting an adversarial signal requires solving a costly optimization problem[10].
   b. *synthesis-based:* these approaches synthesize a new signal for every given test sample using a generative model, particularly the generative adversarial network (GAN) (Samangouei, Kabkab & Chellappa, 2018b). These defense algorithms do not directly filter input signals, and in fact, they find a safe vector for the GAN to craft a signal similar to the original samples. However, not only these defense approaches often suffer from instability[11] and mode collapse[12] issues during GAN training, but also finding the safe vector is usually computationally expensive.

These two approaches are able to make a trade-off between maximizing the recognition accuracy of the algorithm, minimizing the success rate (fooling rate) of the attacks, and

[10] For instance, this might take on average 48 seconds for a six-seconds-length speech signal with the sampling rate of 8 kHz on a NVIDIA GTX-1080-Ti-11 GB memory.

[11] Such as exploding weight vectors for either generator or discriminator especially at larger iterations. Generating oversmoothed samples is a common consequence of instability during training. We refer to Chapter 2 for more information.

[12] Losing sample variations and memorizing a limited number of modes.

managing the computational complexity towards finding the aforementioned safe vector. A reliable defense approach should make this trade-off as optimal as possible so that it works in real-time.

3. *Building Stronger Models:* although it has been proven that adversarial examples exist for all data-driven models on any scales (Papernot, 2018), it is possible to reconfigure the architecture of the classifiers to:
   a. relatively reduce the fooling rate of the attacks,
   b. or at least maximize the cost of the attack for the adversary.

   Moreover, it is feasible to implement some basic settings during data preparation to limit the performance of the attack optimization procedure. For instance, we can find an optimal sampling rate in converting a signal into a spectrogram in order to increase the cost of the attack. Compared to the first and second conditions mentioned above, building a stronger model is very subjective and it might require many try-and-error operations. However, unlike those two conditions, it does not add any computational overhead during training.

This thesis explores possible approaches that somehow contribute to developing algorithms for defending the cutting-edge ESC and ASR models subject to at least one of the three conditions mentioned above.

1. *Better Recognition Algorithm:* focusing on devising a reliable adversarial defense approach, does not diminish the importance of developing algorithms with higher recognition accuracies. More specifically, toward designing a defense approach that meets the conditions mentioned above (especially the first and the third), we need to balance the recognition accuracy of the classifier carefully. For instance, this is possible through developing new data augmentation schemes and/or more complex algorithm architectures.

2. *Faster and more Robust Adversarial Attack*[13]: this explicitly contributes to reducing the computational complexity of the defense algorithm in adversarially training frameworks. Additionally, it helps to understand the functionality of adversarial attacks and the victim models in greater detail.

Research Objectives and Contributions

The main objective of this thesis is to develop approaches for defending data-driven ESC and ASR models against varieties of adversarial attacks with both 1D and 2D representation domains. Although we oblige all the approaches to comply with at least one of the reliability conditions discussed in the previous subsection[14], developing a fully reliable defense algorithm is still an open problem. The most significant advantage of imposing those conditions is to avoid introducing defense algorithms that provide false senses of security against strong adversarial attacks.

We first study a large collection of classifiers proposed for ESC from conventional (e.g., spherical k-means) to advanced deep learning-based classifiers. Upon conducting this investigation we discuss that providing a more comprehensive sample distribution outweighs designing complex architectures for the classifier.

Then, we propose an augmentation technique based on a high-level feature transformation using a cycle-consistent GAN. This technique considerably improves the recognition accuracy of the benchmarking classifiers, in addition to our proposed novel architecture. However, training the GAN in this framework might not result in a stable model. Regarding this concern, in our second step, we develop another classification algorithm to make a better trade-off between high recognition accuracy and low attack success rate. Moreover, we replace a costly high-

[13] Robust in terms of preserving the optimized adversarial perturbation after consecutive playbacks over the air. We discuss this aspect in Appendix V.

[14] In the problem statement section.

level augmentation algorithm with a low-level transformation and spectrogram reconstruction operations to enhance both the quality and quantity of the signals.

Third, towards developing stronger classifiers, we extend the latter work to study the relation between fooling rate of the attack algorithm and signal representation[15]. We demonstrate that it is possible to achieve a more reliable model without implementing any spectrogram reconstruction operation. Nevertheless, this approach fits well for ESC models, which have been trained on 2D representations. For covering 1D signals, in our fourth step, we introduce a novel defense approach against adversarial attacks developed for end-to-end ASR systems.

Since this thesis is manuscript-based, each chapter presents a different journal publication about developing defense algorithms against a comprehensive list of cutting-edge adversarial attacks. Overall, there are nine major contributions in this work which led to four journals and five conference publications[16]. These contributions are listed below:

1. Introducing an unsupervised ESC algorithm using random forests (RF). We show that training this algorithm on benchmarking datasets augmented with a cycle-consistent GAN outperforms a few advanced deep learning classifiers (e.g., GoogLeNet and AlexNet). This GAN employs a sequence of residual-convolutional architectures separately for the generator and discriminator networks to provide a wider range of distinguishable features to the front-end classifier. Our focus in this paper is improving the recognition accuracy of the classifier over a baseline and state-of-the-art models. Our underlying motivation[17] for selecting RF rather than other data-driven configurations was its higher resiliency against targeted adversarial attacks, specifically relative to deep learning architectures.

   This publication had a significant role in getting us into the right direction towards securing ESC and ASR classifiers against adversarial signals. Furthermore, our conducted

[15] In fact, this study investigates the effect of spectrogram production settings on the model robustness.

[16] Excluding two additional papers which are already published without peer-review, i.e. on arXiv.org.

[17] Inspired by Papernot (2018).

experiments on this ensemble-based algorithm led us to properly define the reliability conditions as stated in the previous subsection[18].

Related publication:

a. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2019). "Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network." Elsevier Applied Soft Computing, 86, 105912.

2. Our second contribution is proposing an adversarial defense algorithm respecting all of our predefined reliability conditions. We demonstrate that this approach makes a reasonable trade-off between model accuracy and robustness against a wide range of adversarial attacks. Finally, in order to better analyze the performance of our defense approach using a supervised classification algorithm, we cross-examine its robustness against two attack groups:
   a. algorithms which have been primarily developed for attacking conventional classifiers (e.g., support vector machines - SVM),
   b. algorithms designed for attacking deep-leaning-based architectures (e.g., variants of convolutional neural networks - CNN).

   Although these two groups are fundamentally different, they both implement optimization-based procedures to fool the advanced data-driven recognition models toward any incorrect phrases. As a part of our experiments during the development of this defense approach, we also investigate adversarial signal transferability among conventional and deep learning-based ESC models. Furthermore, for comparing the robustness of our proposed supervised classification algorithm with other models, we measure their adversarial resiliency score using the local intrinsic dimensionality (LID) metric (Ma, Li, Wang, Erfani, Wijewickrema, Schoenebeck, Song, Houle & Bailey, 2018). Finally, inspired by LID, we develop another criterion for detecting adversarial signals using a subspace measurement technique.

[18] Inside the problem statement subsection.

Related publications:

- Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2019). "A robust approach for securing audio classification against adversarial attacks." IEEE Transactions on Information Forensics and Security (TIFS), 15, 2147-2159.
- Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2020). "Detection of adversarial attacks and characterization of adversarial subspace." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020), pp. 3097-3101.

3. Our third contribution is characterizing the relation between signal representation (i.e., MFCC, STFT, and DWT) and the model's recognition accuracy as well as the fooling rate of the attacks. This publication primarily concerns our third reliability condition, which obliges us to develop an attack-resilient[19] classifier from the base instead of devising a separate defense algorithm on top of the achieved recognition model. We identify primary spectrogram settings upon carrying out extensive experiments, which considerably affect the cost of attack (the number of required gradient computations) for the adversary averaged over the allocated budgets. Thus, from a statistical point of view, this algorithm might not constitute an upscale defense approach. However, it finds a way to maximize the cost of attack for the adversary.

   In connection with our second contribution (Esmaeilpour, Cardinal & Koerich, 2020), we also show that the MFCC representation has a relatively lower adversarial transferability ratio among advanced deep learning architectures. Therefore, this defense approach also meets the first reliability condition, which aims to avoid gradient obfuscation. In order to investigate the satisfaction of the second reliability condition, we benchmark this algorithm into the adversarially training framework. This framework helps to measure the robustness of this defense approach subject to making a balance between recognition accuracy and the total number of required gradient computations.

[19] At least partially resilient.

Related publications:

a. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2020). "From sound representation to model robustness." Currently under review at Elsevier Applied Acoustics Journal.

b. Sallo, R. A., Esmaeilpour, M., and Cardinal, P. (2020-2021). "Adversarially Training for Audio Classifiers." In 25th International Conference on Pattern Recognition (ICPR), (pp. 9569-9576). IEEE.

4. Our fourth contribution is developing a novel algorithm for defending ASR systems, from conventional speech-to-text models (e.g., Kaldi) to modern LSTM-based neural network architectures (e.g., DeepSpeech) against varieties of adversarial attacks. This defense algorithm complies with our predefined reliability conditions, specifically the first two items. Our main concentration in this publication is defending end-to-end systems against adversarial attacks for safe real-time transcription. However, all the findings and experiments are bijectively generalizable to spectrogram-based classifiers.

   Our proposed algorithm is based on implementing a multi-discriminator GAN defined in the restricted Sobolev space (Brezis, 2010). Since this defense approach's performance is deeply dependent on the generalizability of the GAN, we also propose a new regularization technique for stable training. In fact, this technique is an extension of our previous publication about adversarial detection (Esmaeilpour *et al.*, 2020b). Moreover, we introduce simple yet effective architectures for both the generator and discriminator networks towards smoothly training the GAN.

   One of our potential concerns about this defense approach is occurring mode collapse (Mao, Li, Xie, Lau, Wang & Smolley, 2018) during training the GAN at later iterations. This might pose a security issue on the robustness of this algorithm. In response to this concern, we develop two additional defense approaches:

   a. introducing a novel configuration for the GAN based on the class-conditional generative model (Mirza & Osindero, 2014),

b. proposing a cyclic GAN including two generator networks for regularizing the joint learning curve.

Since these two algorithms are synthesis-based, they meet the reliability conditions mentioned above for adversarial defense. However, they might be computationally expensive especially for very long[20] and multi-speaker speech signals. This motivated us to pursue our research toward developing a fast adversarial attack and using the crafted signals for adversarially training the ASR algorithm. Toward this end, we introduce a novel attack algorithm using the Cramér integral probability metric (CIPM) (Bellemare, Danihelka, Dabney, Mohamed, Lakshminarayanan, Hoyer & Munos, 2017). However in this thesis, we neither discuss nor analyze the implementation of adversarially training for our CIPM-based attack.

Related publications:

- Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2021). "Multi-Discriminator Sobolev Defense-GAN Against Adversarial Attacks for End-to-End Speech Systems." Currently under review at IEEE Transactions on Information Forensics and Security (TIFS) Journal.
- Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2021). "Class-Conditional Defense GAN Against End-to-End Speech Attacks." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2021), pp. 2565-2569.
- M. Esmaeilpour, P. Cardinal and A. L. Koerich, "Cyclic Defense GAN Against Speech Adversarial Attacks," in IEEE Signal Processing Letters, vol. 28, pp. 1769-1773, 2021, doi: 10.1109/LSP.2021.3106239.
- Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2021). "Towards Robust Speech-to-Text Adversarial Attack". Currently under review at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2022).

[20] On average, above 140 seconds per sequence.

Organization of the Thesis

This is a manuscript-based thesis. Thus every chapter corresponds to a journal paper. Furthermore, papers that have been published in conference venues are organized into appendices. Figure 0.1 depicts the organization of this thesis through a flowchart. As shown, journal and conference publications are illustrated by blue and green boxes, respectively. Solid and dotted arrows outline the type of relation among chapters and appendices. A solid arrow specifies that a source publication should be read before its associated target publication to better understand the discussed concepts and algorithms. Similarly, a dotted arrow designates a suggested reading among chapters and appendices.

This thesis starts with an overview of the state-of-the-art ESC and ASR models designed for real-time transcriptions (Chapter 1). We briefly review the properties of these models and characterize the threat of adversarial attacks for them. Moreover, we summarize all the introduced approaches for defending these models (both end-to-end and representation-level) against varieties of adversarial attacks.

Chapter 2 presents an unsupervised feature learning approach for ESC using a weighted cycle-consistent GAN. In this chapter, we review all the cutting-edge algorithms developed for ESC and introduce our classification approach focusing on two major aspects:

1. high-level data augmentation for improving recognition accuracy of the classifiers. We develop a novel residual-convolutional GAN architecture for such an aim and demonstrate its superior performance over low-level augmentation techniques;
2. designing an unsupervised ensemble-based front-end classifier to increase the chance of resiliency against targeted and non-targeted adversarial attacks. Although we do not directly discuss adversarial defense techniques in this chapter, the proposed classifier meets our predefined reliability conditions.

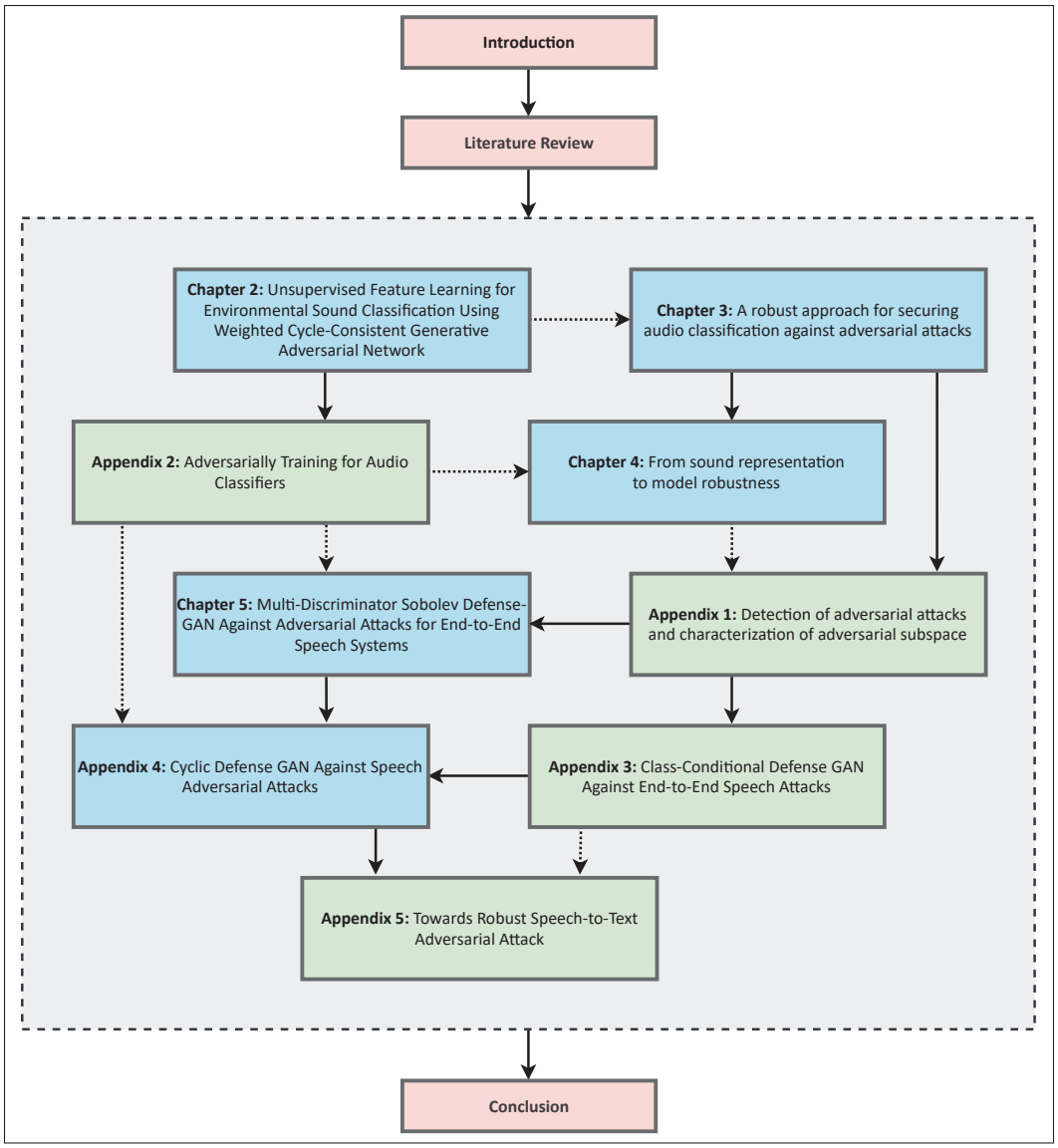

Figure 0.1 Overview of the thesis chapters. Journal and conference publications are shown in blue and green boxes, respectively. Additionally, solid arrows indicate the flow of dependency among chapters and appendices (i.e., the source should be read before the associated target). Likewise, the suggested readings which contribute to better understanding the concepts of the chapters and appendices are shown in dotted arrows.

Chapter 3 presents a robust approach for securing ESC models against white and black-box adversarial attacks. We argue that it is possible to make a reasonable trade-off between the victim

model's recognition accuracy and fooling rate. More specifically, we improve this trade-off by proposing a new classification algorithm and few signal augmentation techniques: color compensation for spectrograms, highboost filtering, dimensionality reduction and convolutional denoising autoencoders. Additionally, we investigate the transferability of adversarial signals bijectively from attacks that have been developed against conventional and deep learning-based architectures.

Chapter 4 presents a summary of our extensive experiments on identifying the relation between spectrogram production settings (e.g., sampling rate, frame length, feature normalization, etc.) and the model's fooling rate against adversarial attacks. These experiments have been conducted for addressing the third item of our predefined defense reliability conditions, which obliges us to design more robust recognition algorithms. The front-end classifier in this chapter is a 16-layers residual-based CNN without running any costly high-level data augmentation procedures. The motivation behind using this classifier is its superior performance compared to other CNN architectures both in terms of higher recognition accuracy and a relatively lower number of required training parameters.

In Chapter 5, we introduce a novel adversarial defense approach according to our predefined reliability conditions for end-to-end ASR systems. This algorithm is based on synthesizing a new speech signal for every given test input. The main constraint during this synthesis procedure is crafting a naturally-sounding signal relative to the available original recordings. Toward this end, we develop a multi-discriminator GAN implemented in the restricted Sobolev space. Our primary motivation for choosing this space is its strong correlation with the Fourier transform coefficients such as matrices of MFCC and STFT spectrograms. We experimentally prove that our proposed defense approach outperforms other algorithms both in terms of attack success rate and preserving the quality of the signals.

Appendix I characterizes the possibility of statistically measuring the distance between adversarial and original signal subspaces using the chordal distance metric computed in the Schur decomposition domain (Van Loan & Golub, 1983). This metric achieves a small value between signals which lie in the same subspace and a large value for samples in dissimilar subspaces. Using this metric, we experimentally prove that adversarial signals are conveniently distinguishable from original and noisy samples. As shown in Figure 0.1, this paper is an essential prerequisite for correctly understanding the defense approach proposed in Chapter 5.

Appendix II investigates the effect of adversarially training ESC models to address our second defense reliability condition. We conduct our experiments on a wide range of complex classifiers including modern deep learning blocks such as inception, attention, convolution, and residual. We show that adversarially training considerably reduces the performance of the classifier. However, it improves its robustness against six types of targeted and non-targeted adversarial signals by constraining the attack algorithm over the maximum required adversarial perturbation to a specific threshold.

Appendix III proposes a novel defense algorithm for defending both conventional and modern end-to-end ASR systems. This approach is also based on synthesizing a signal seamless to any given test speech recording. Moreover, it does not obfuscate gradient information[21]. Thus it provides a correct sense of security for these transcription systems. The major difference between this algorithm and the defense approach proposed in Chapter Chapter 5 is two-fold:

1. it introduces a class-conditional configuration with multiple embeddings instead of Sobolev-GAN,
2. it implements simpler architectures for both the generator and discriminator networks.

[21] This refers to some operations which manipulate the gradient vectors and consequently provide false sense of security. More information can be found in (Athalye *et al.*, 2018b).

However, this algorithm partially degrades the quality of the signals during the synthesis procedure. We address this concern as the following.

In Appendix IV, we introduce our third synthesis-based defense approach for end-to-end transcription systems. Although herein we do not discuss the generalization of this algorithm to the representation-level, our initial implementation corroborates its usefulness for models trained on spectrograms. The generative model used in this approach is the standard least-square GAN (Hong, Hwang, Yoo & Yoon, 2019), but with a new architecture. Therefore, reading Appendix I is an essential prerequisite for understanding this defense approach.

Appendix V introduces a novel adversarial attack algorithm primarily for ASR systems. The major motivation for developing this attack is three-fold:

1. proposing an algorithm for attacking super-complex speech-to-text models to work in real-time[22],
2. since crafting adversarial signals is computationally expensive, developing a fast attack might improve the adversarially training procedure,
3. enhancing the resiliency of adversarial signals after playbacks over the air.

Summary

In this chapter, we explained the subject of this thesis and defined three reliability conditions for securing ESC and ASR models against varieties of adversarial attacks. Moreover, we mentioned our objectives and listed our contributions that have been published in journal and conference venues. In the next chapter, we provide a comprehensive literature review on defense algorithms developed for securing both ESC and ASR models against adversarial attacks.

[22] At least partially.

CHAPTER 1

LITERATURE REVIEW

This chapter provides a detailed survey and comprehensive analysis of the state-of-the-art adversarial defense algorithms developed for securing both end-to-end and representation-level ESC and ASR systems. We start this chapter by concisely explaining signal representation procedures that have become standard signal processing techniques over the last few decades. Then, we briefly review the history of data-driven classification for audio and speech recordings. Following this analysis, we characterize the existence of adversarial attacks for these models. Finally, we provide the taxonomy of adversarial defense approaches with a brief discussion over the imposed challenges.

1.1 Signal Representation

In this section, we explain three standard signal representation techniques, which have been essentially developed for extracting more informative features from a 1D audio or speech recording. In a big picture, there are three main motivations for developing signal representation: (Pickett, 1999; Gold, Morgan & Ellis, 2011):

1. Acquisition devices such as standard microphones continuously record signals with considerable overlap over time to capture a smooth signal. This operation contributes to audiotape a high-quality signal somewhat without noticeable artifacts. However, it increases the dimensionality of the signal;
2. As stated above, raw 1D signals have high dimensionality and training a data-driven classifier on them often requires training more parameters. Generally, a signal's dimensionality is highly dependent on its sampling rate, duration, and number of channels. For instance, a one-second-length signal with the minimum sampling rate of 8kHz distributed over two channels contains 16k data-points. Obviously, generating a low-dimensionality representation for such a signal improves the computational complexity of the learning algorithm considerably;

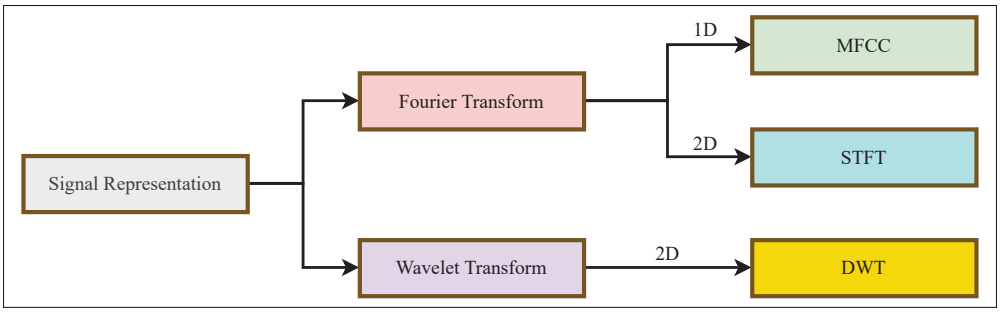

Figure 1.1 General taxonomy of signal representation for ESC and ASR

3. Environmental noises, including microphone-speaker reverberation, room impulse response, and echo, have always been among the damaging side-effects of digital recording (Sterne, 2015). Converting a recorded signal into a frequency representation substantially reduces such side effects.

Generally, there are two main transformations for generating signal representation (spectrograms), namely Fourier and wavelet. Figure 1.1 shows the general taxonomy of representations according to their associated transformations. In the following subsections, we briefly explain the theories behind each of these mainstream spectrograms.

1.1.1 Short-Time Fourier Transform: STFT

For a given single or multichannel continuous signal $a(t)$, which is distributed over time (t), we compute its STFT representation as (Pickett, 1999):

$$\text{STFT}\left\{a(t)\right\}(\tau,\omega) = \int_{-\infty}^{\infty} a(t)w(t-\tau)e^{-j\omega t}dt \tag{1.1}$$

where $w(\cdot)$ is a Hann window function for smoothing purposes. Additionally, ω and τ are frequency and time axes, respectively. This transform is fully generalizable to discrete-time domain with any number of channels, as well.

Assume a discrete signal $a[n]$ distributed over n equidistance digital components (discrete time) is given. The STFT representation for this signal can be computed as (Pickett, 1999):

$$\text{STFT}\left\{a[n]\right\}[m,\omega] = \sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n} \tag{1.2}$$

where $m \ll n$ and ω is a discrete frequency coefficient. For capturing more local features from $a[n]$, it is conventionally recommended to divide it into overlapping sub-signals and compute Fourier transform on every resulting chunk (Pickett, 1999). This results in achieving an array of complex coefficients for the entire signal. Computing the power spectrum of this array yields a 2D STFT spectrogram as follows.

$$\text{Sp}_{\text{STFT}}\left\{a[n]\right\}[m,\omega] = \left|\sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n}\right|^2 \tag{1.3}$$

This representation plots the frequency distribution of the given signal over discrete-time and compared to both original signal $a(t)$ and $a[n]$, it has much lower dimensions. Nevertheless, this transform is a lossy operation and we might lose some information during the transformation (Pickett, 1999).

1.1.2 Mel-Frequency Cepstral Coefficient: MFCC

This transform is a condensed variation of the STFT with a few additional postprocessing operations including nonlinear transformation and orthogonal normalization (Pickett, 1999). After STFT production, we multiply every column of the achieved spectrogram with a number of predefined Mel-filter banks (power estimates for amplitudes distributed over discretized frequency components). This operation yields complex vectors with correlated distributions. For enhancing the resolution and quality of the resulting vectors, it is recommended to run a logarithmic filtration procedure (Pickett, 1999). The last step is mapping these filtered vectors into another 1D representation using the discrete cosine transform (DCT). The motivation behind employing DCT is decorrelating vectors with similarities above a predefined threshold and

consequently reducing dimension of the input signals (Pickett, 1999; Han, Chan, Choy & Pun, 2006; Zhu & Alwan, 2001).

During the last few decades, this representation has been widely used for audio and speech enhancement and more importantly, for classification purposes. The majority of ASR models use MFCC representation such as Kaldi and the cutting-edge DeepSpeech systems. Furthermore, this 1D spectrogram has been well established as a fairly standard representation for conventional generative models employing Markov chain and Gaussian mixture models (Shi, Ahmad, He & Chang, 2018; Maurya, Kumar & Agarwal, 2018).

1.1.3 Discrete Wavelet Transform: DWT

In terms of functionality, the DWT is almost the same as STFT. However, the latter is simpler since it uses more straightforward basis functions[23]. Discussion on the advantage of either of these two transformations over another is out of the scope of this thesis. However, we mention a few of their differences (Chui, 1993; Pickett, 1999; Jensen & la Cour-Harbo, 2001; Young, 2012; Addison, 2017):

1. the STFT only provides a spectrum of frequency distributions (from low to high), but DWT provides a similar spectrum with the localization of frequency components,
2. basis functions in DWT (a.k.a. mother functions) are not limited to the perpendicular $\sin(\cdot)$ and $\cos(\cdot)$, and they include a variety of complex mathematical relations,
3. correlation of frequency components in DWT is usually greater than STFT. This sometimes interprets as an advantage and is often considered as a disadvantage. In fact, the correct interpretation is relative to the application.

[23] The functions used for bijectively converting $a[n]$ to frequency representation. For instance, the basis functions used for the STFT and its variants are $\sin(\cdot)$ and $\cos(\cdot)$.

Mathematically, the DWT maps the continuous signal $a(t)$ into time and scale (frequency) coefficients through convolving it with a predefined basis function (Chui, 1993):

$$\text{DWT}\left\{a(t)\right\} = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} a(t)\psi\left(\tfrac{t-\tau}{s}\right) dt \tag{1.4}$$

where s and τ denote the scale and time variations, respectively. Additionally, ψ is the mother function (the core basis function as defined in Eq. 1.5). There are a variety of mother functions for different applications and to the best of our knowledge, there is no analytical way to find the most optimal function. However, the complex Morlet has always been among common mother functions especially for ESC and ASR applications (Stephane, 1999). The formal definition for this function is as follows (Stephane, 1999):

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega t} e^{-t^2/2} \tag{1.5}$$

where the scale of t^2 is subjective and it can be changed during convolution with $a(t)$. Eq. 1.4 can be straightforwardly generalized to discrete signals such as $a[n]$ without any additional computational overhead (see Eq. 1.6).

$$\text{DWT}\left\{a[k,n]\right\} = \sum_{n=-\infty}^{\infty} a[n]\psi\left[n,k\right] \tag{1.6}$$

where n and k are integer values for the discrete mother function. The power spectrum for this transformed signal is a 2D array:

$$\text{Sp}_{\text{DWT}}\left\{a[n]\right\} = \left|\text{DWT}\left\{a[k,n]\right\}\right|^2 \tag{1.7}$$

where it yields a 2D spectrogram for any given audio or speech signal. In the following sections, we review the state-of-the-art machine learning algorithms designed for dealing with representation-level and 1D signals.

1.2 Environmental Sound Classification: ESC

Regarding the vast applications of ESC for real-life challenges, especially for scene understanding (Kim, Sundaram, Georgiou & Narayanan, 2009; Brust, Sickert, Simon, Rodner & Denzler, 2015), numerous algorithms have been introduced thus far. In general, we can organize the proposed approaches into different categories as the following.

1.2.1 Shallow Vs. Deep Learning-Based Classifiers

Conventional classifiers refer to all the non-deep neural network-based data-driven algorithms, which usually use hand-crafted features (derivable from 1D signals such as MFCC or spectrograms) for training and evaluation purposes, such as SVM (Wang, Wang, He & Hsu, 2006; Valero & Alías, 2012a,b; Umapathy, Krishnan & Rao, 2007), nearest neighbour (da Silva, W Happi, Braeken & Touhafi, 2019; Tsalera, Papadakis & Samarakou, 2020), random forest (Piczak, 2015b; Homsi, Medina, Hernandez, Quintero, Perpiñan, Quintana & Warrick, 2016), hidden Markov model (Vacher, Serignat & Chaillol, 2007; Su, Yang, Lu & Wang, 2011; Ling-li, 2011), adaptive boosting (Chiu, Gestner & Anderson, 2011), bootstrap aggregation (Alsouda, Pllana & Kurti, 2019), cascading (Foggia, Saggese, Strisciuglio & Vento, 2014), classification tree (Breiman, Friedman, Olshen & Stone, 1984), learning vector quantization (Tang, Liu, Chen, Zhou & Ding, 2007; Syafria, Buono & Silalahi, 2014), extreme learning machine (Ahmad, Agrawal, Joshi, Taran, Bajaj, Demir & Sengur, 2020), etc. These algorithms, especially SVM-based approaches, have been very popular before the proliferation of deep learning configurations. The major challenge in all the abovementioned algorithms is handcrafting a comprehensive set of features from either 1D signals or the associated spectrograms (Hu, Xu & Wu, 2007).

During the last decade, with the remarkable progress in both software and particularly hardware developments, data-driven classifiers have been mostly focused and redirected towards deep learning architectures. These architectures bypass the challenging part of feature extraction in a conventional classification framework, and they relatively provide a wider learning subspace. To name a few of these algorithms, we can mention variants of convolution (Zhang, Zou & Shi,

2017; Salamon & Bello, 2017; Sailor, Agrawal & Patil, 2017; Su, Zhang, Wang & Madani, 2019; Zhang, Zou & Wang, 2018a; Park & Yoo, 2020; Chen, Guo, Liang, Wang & Qian, 2019; Mushtaq & Su, 2020; Park & Yoo, 2020; Lu, Ma, Liu & Qin, 2021), recurrent (Guzhov, Raue, Hees & Dengel, 2021; Palanisamy, Singhania & Yao, 2020; Nasiri & Hu, 2021), attention (Zhang, Xu, Zhang, Qiao & Cao, 2019b; Sharma, Granmo & Goodwin, 2019; Miyazaki, Komatsu, Hayashi, Watanabe, Toda & Takeda, 2020), and LSTM architectures (Rahman, Rahman, Hossain, Hossain, Akhond & Hossain, 2021; Constantinou, Michaelides, Alexopoulos, Pieri, Neophytou, Kyriakides, Abdi, Reodica & Hayes, 2021; Chandrakala *et al.*, 2021). Compared to conventional algorithms, deep learning-based classifiers often yield a more accurate (mainly in terms of generalizability) model at the cost of requiring much more training parameters (Zhang, Quan & Ren, 2016).

In addition to the categories mentioned earlier, some hybrid ESC algorithms use the combination of conventional and deep learning architectures (Agrawal, Sailor, Soni & Patil, 2017; Demir, Turkoglu, Aslan & Sengur, 2020; Akbal, 2020). In most of these algorithms, the employed neural network replaces the traditional feature extraction techniques to provide more discriminative feature vectors to the front-end conventional classifier. There are many debates on developing hybrid algorithms. However, we do not discuss them in this thesis. Instead, we mention a few of our own experiences upon conducting exploratory experiments on hybrid algorithms:

1. they usually cannot outperform dense CNNs and the associated variants (e.g., residual-convolutional networks),
2. they often require fewer training parameters compared to fully deep learning-based classifiers,
3. they unanimously make a better trade-off between model's recognition accuracy and the adversarial attack robustness[24].

1.2.2 End-to-end Vs. Representation-Level Learning

The end-to-end classification framework includes all the algorithms which do not use any types of representations for training, neither 1D nor 2D spectrograms. In terms of cardinality, the

[24] We refer to Chapter 2 and 3 of this thesis.

total number of representation-level classifiers is much more than end-to-end counterparts. Furthermore, the highest recognition accuracy has often been reported for algorithms trained on spectrograms (Boddapati *et al.*, 2017). However, this does not diminish the importance of developing end-to-end ESC models since:

1. unlike all the representation-level algorithms, the end-to-end classifiers use the entire information of the signal[25],
2. developing an adversarial signal is much more costly than developing an adversarial spectrogram. Technically, this does not constitute a defense policy. However, it makes the attack more costly for the adversary.

EnvNet (Tokozume & Harada, 2017) and EnvNet-V2 (Tokozume, Ushiku & Harada, 2018) are among the latest ESC models which introduce a new perspective in end-to-end learning. More specifically, they employ the between-class learning policy for effective training. This policy contributes to maximizing intra-class similarity and minimizing over the inter-class correlation. Another novel end-to-end algorithm has been introduced in line with this approach, which implements a multiresolution CNN (Zhu, Xu, Wang, Zhang, Li & Peng, 2018b). All these 1D algorithms have been successfully evaluated on standard benchmarking environmental sound datasets such as ESC-10, ESC-50 (Piczak, 2015b), and UrbanSound8K (Salamon, Jacoby & Bello, 2014a).

The majority of the latest recognition algorithms which fit in the representation-level category are based on dense CNNs such as variants of Piczak-CNN (Piczak, 2015a; Salamon & Bello, 2017; Agrawal *et al.*, 2017; Tak, Agrawal & Patil, 2017). All these networks employ different architectures and they have been designed to improve both the performance and the model's generalizability. VGG-like architectures are also among popular configurations for ESC (Zhang *et al.*, 2019b; Zhang, Xu, Cao & Zhang, 2018b). This is presumably due to the popularity of VGG networks in the computer vision domain (Simonyan & Zisserman, 2015). All these algorithms have also been experimented on challenging datasets and they have demonstrated recognition

[25] Converting a signal into a spectrogram is considered a lossy operation. Thus, representation-level algorithms might not have access to the entire information of the designated signals.

accuracies competitive to the human level of understanding (Wang, Zou, Chong & Wang, 2019; Boddapati *et al.*, 2017).

1.2.3 Semi-supervised Vs. Supervised Learning

In general, we can also categorize all the introduced ESC algorithms into semi-supervised and supervised groups since unsupervised approaches take up a small portion of the data-driven recognition models. We acknowledge the existence of many debates in this regard. However, we do not cover them herein since they are out of the scope of this thesis.

Supervised algorithms are fairly straightforward and we have mentioned the majority of them in Sections 1.2.1 and 1.2.2. Therefore, herein we briefly review some cutting-edge semi-supervised algorithms.

For bridging the gap between the scarcity of labeled data and the optimal amount of training signals in the context of ESC, many approaches have been introduced during the last decade (Han, Coutinho, Ruan, Li, Schuller, Yu & Zhu, 2016; Zhang & Schuller, 2012; Serizel, Turpault, Eghbal-Zadeh & Shah, 2018; Bodini, 2019). These algorithms have been proposed to developing high-resolution models with a limited number of training signals, such as the ESC-10 dataset (Piczak, 2015b), which contains only 2k five-second-length environmental sounds. Competitive to these approaches, an echo state networks-based semi-supervised algorithm has been proposed to tackle small datasets' training issues (Scardapane & Uncini, 2017). Moreover, semi-supervised classifiers have been extensively exploited for audio tagging and event detection (Chu, Narayanan & Kuo, 2009b; Akiyama & Sato, 2019; Cances & Pellegrini, 2021; Lin, Wang, Liu & Qian, 2020).

In the next section, we review the state-of-the-art algorithms developed for ASR applications, particularly for real-time speech-to-text transcription.

1.3 Automatic Speech Recognition - ASR: Speech-to-Text Transcription

Over the last decades, impressive progress has been made in developing automatic speech-to-text transcription systems. Nowadays, these advanced systems are among the inevitable services of all the smart devices, from personal digital assistances (e.g., cellphones, iPad, etc.) to voice command devices in modern vehicles, especially autonomous cars. In this section, we only review such state-of-the-art systems, which have been:

1. designed to work in adverse scenarios efficiently[26],
2. benchmarked for developing adversarial attack and defense algorithms.

Therefore, other aspects of ASR systems such as speaker identification, linguistic modeling, language identification, speech coding, acoustic-phonetic simulations, multimodal processing, emotion recognition, etc. do not fit this thesis's context. In the following, we briefly study three main benchmarking speech-to-text systems.

1.3.1 Kaldi

This is one of the popular speech-to-text transcription systems which has been properly maintained over the last few years. This system has been gradually adapted to develop the deep learning algorithms initiated by Povey *et al.* (2011). Kaldi contains four major processing layers as follows[27].

1.3.1.1 Extracting Acoustic Features

The first layer extracts features from the given speech signals since Kaldi is a representation-level ASR model. The main representation scheme used in this speech-to-text system is MFCC. However, it also uses the perceptual linear prediction (PLP) feature extraction technique (Hönig, Stemmer, Hacker & Brugnara, 2005). PLP is a variant of STFT that employs loudness adjustment

[26] With the existence of surrounding noises and environmental sounds.

[27] We are aware that Kaldi includes additional processing layers; however we have intentionally excluded them so as to avoid presenting unnecessary complicated concepts.

and recursive cepstrum computation procedures (Childers, Skinner & Kemerait, 1977) to achieve more comprehensive signal descriptors[28].

1.3.1.2 Training Monophone and Triphone Models

This acoustic model does not contain any explicit contextual information about phonetics (neither the preceding nor the subsequent phonetics). Phonetic refers to any distinction of the given speech signal, and in fact, every signal contains a chronological series of analytical phonetics[29]. The training procedure can be implemented with the hidden Markov model (HMM), Gaussian mixture model (GMM), or deep neural network (DNN). All these generative models yield a vector of embeddings associated with every monophone[30]. The earlier versions of Kaldi used to work with HMM plus GMM embeddings, and recently they have been augmented (replaced) with DNNs.

1.3.1.3 Aligning Signals with the Achieved Acoustic Models

This layer (a.k.a. Viterbi training) is for tuning the achieved embeddings with the ground-truth references (Franzini, Lee & Waibel, 1990; Spitkovsky, Alshawi, Jurafsky & Manning, 2010). The motivation behind taking this step is that all the obtained parameters upon training the aforementioned acoustic modelings may not necessarily represent an accurate sample distribution. Therefore, it is recommended to cycle through the layers of acoustic modeling and alignment iteratively.

Viterbi training is an important process in Kaldi's speech-to-text transcription, especially for utterances including multispeaker signals (Yoma & Villar, 2002). This realignment procedure

---

[28] Feature vector(s).

[29] We refer to Rabiner (1989) for more explanations about phonetics (a.k.a. phone or embedding) modeling.

[30] Monophone is a single phonetic element, and triphone denotes a phoneme variants with the presence of left and right phonemes.

directly contributes to correctly tuning the front-end classifier, as well (Arik, Diamos, Gibiansky, Miller, Peng, Ping, Raiman & Zhou, 2017).

1.3.1.4 Linear Discriminant Analysis: LDA

This layer prepares the HMM states using LDA and maximum likelihood linear transform (Leggetter & Woodland, 1995; Gales, 1998) for the front-end classifier. Moreover, these two operations reduce the feature space for yielding a unique transformation for every speaker. The last step is adaptive training (Anastasakos, McDonough, Schwartz & Makhoul, 1996) to map embeddings into tokens[31].

In this subsection, we briefly reviewed the Kaldi speech-to-text system, which uses a combination of HMM, GMM, and DNN configurations. Recently, it has been demonstrated that fully DNN-based transcription models can outperform conventional systems without Kaldi's time-consuming preprocessing operations (Mozilla-DeepSpeech, 2017). We review two of such systems in the following subsections.

1.3.2 DeepSpeech

This speech-to-text transcription system is based on the advanced recurrent neural network (RNN) and similar to Kaldi, it is also categorized into the representation-level speech recognition model. The type of spectrogram used in this RNN-based system is MFCC, and since there are many variants for DeepSpeech (Hannun, Case, Casper, Catanzaro, Diamos, Elsen, Prenger, Satheesh, Sengupta, Coates et al., 2014), herein, we only focus on Mozilla's standard implementation (Mozilla-DeepSpeech, 2017).

Mozilla's DeepSpeech exploits sequences of stacked LSTMs with five layers of hidden units. The last layer of this architecture is bidirectional (Huang, Xu & Yu, 2015), followed by forward and backward recurrence modules before the softmax operation. This configuration contributes

[31] Characters without duplication. Herein, additional postprocessing operations are excluded for clarity purposes.

to having a larger memory capacity distributed over time and it also better captures local distributions of the MFCC representation. Moreover, this system employs the standard jittering regularization scheme (Krizhevsky, Sutskever & Hinton, 2012) for convenient sequence-to-sequence translations. As a result, DeepSpeech has been efficiently implemented and fully supports parallel computation for real-time transcription. The main advantages of DeepSpeech over Kaldi is twofold:

1. unlike Kaldi, it does not incorporate any costly preprocessing, acoustic modeling, and the iterative embedding tuning procedures,
2. the maintenance and development of DeepSpeech are very straightforward since it is fully implemented in Python. The main core of the standard Kaldi API[32] is written in C++ which may not conveniently cope with some python packages developed for adversarial attacks (e.g., Foolbox(Rauber, Brendel & Bethge, 2017) and Cleverhans (Papernot, Faghri, Carlini, Goodfellow, Feinman, Kurakin, Xie, Sharma, Brown, Roy et al., 2016b)).

The dataset used for training the DeepSpeech is Mozilla common voice[33] (MCV, 2019), including more than 1000 hours of short and long utterances with various dialects, accents, genders, language, and ages. This comprehensive dataset has also been used for transfer learning and the model's fine-tuning purposes among RNN-based architectures (Ardila, Branson, Davis, Henretty, Kohler, Meyer, Morais, Saunders, Tyers & Weber, 2019; Winata, Cahyawijaya, Lin, Liu, Xu & Fung, 2020). Moreover, variants of these recurrent networks with practical APIs have been generalized to other large datasets. In the following, we briefly review one of such strong APIs developed for another huge benchmarking speech dataset.

1.3.3 Lingvo

This speech-to-text API is also a fully deep learning-based transcription system and similar to DeepSpeech, it does not incorporate any costly acoustic modeling procedures (Sutskever, Vinyals & Le, 2014). However, the architecture of this RNN-based system is relatively complex

[32] Application programming interface.

[33] https://voice.mozilla.org/en/datasets

since it uses stacked LSTM layers with attention configurations (Bahdanau, Cho & Bengio, 2015). Lingvo is inspired by Chan, Jaitly, Le & Vinyals (2016) in developing a recognition framework based on the listen, attend, and spell model. This framework launches a dense architecture for sequence-to-sequence modeling using consecutive layers of attention blocks.

Lingvo is also a representation-level transcription system and it primarily uses MFCC for training. However, it adapts well with other representations such as STFT and DWT. This system feeds the achieved MFCC vectors into an encoder which implements a sequence of convolution and recurrent layers. Then, it exploits an LSTM-based decoder to map features into the tokens (characters).

The dataset used for training the Lingvo is LibriSpeech (Panayotov, Chen, Povey & Khudanpur, 2015) which includes a large collection of speech signals with various durations, number of channels, speakers, and environmental settings. There are also smaller variations for this dataset which have also been recognized as standard benchmarking subsets for speech recognition (Zen, Dang, Clark, Zhang, Weiss, Jia, Chen & Wu, 2019). The major concern with LibriSpeech and its variants is that they do not include a comprehensive list of multi-language utterances. Recently, French-speaking utterances have been augmented with the latest release of LibriSpeech, nevertheless transfer learning is still required for model tuning purposes (Kocabiyikoglu, Besacier & Kraif, 2018).

In Sections 1.2 and 1.3 we briefly explained categories of algorithms that have been developed for ESC and ASR. In the following section, we characterize the existence of adversarial signals for these recognition models.

1.4 Characterizing Adversarial Attacks for ESC and ASR Systems

There is no consensus on the initial characterization of adversarial attacks for the data-driven recognition models. However, there have been many publications in this regard since decades ago (Takahata, Imai & Tsuji, 1992; Mustafa, Khan, Hayat, Goecke, Shen, Shao et al., 1995; Xiao, Xiao & Eckert, 2012). This field of research reemerged after the recent development

of deep neural networks with the comprehensive discussion provided by Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow & Fergus (2014). Over the last few years, a large volume of publications has been made and they have seriously challenged the recognition performance of the state-of-the-art classifiers. These attacks have been designed for both end-to-end and representation-level ESC and ASR recognition systems.

Typically, an adversarial attack is an optimization formulation toward crafting inputs for the classification model according to two constraints:

1. they should be seamless to the original samples,
2. they have to redirect the victim model to predict incorrect output(s).

The optimization procedure iteratively finds an optimal perturbation for every original signal subject to fool the model toward wrong predictions. In the following subsections, we explain some properties of adversarial attacks developed against ESC and ASR systems.

1.4.1 Attack Properties

In a big picture, we can describe an adversarial attack algorithm according to its characteristics such as perceptibility of the perturbation, accessibility to the victim model, specificity of the predictions, and metrics for measuring the magnitude of the perturbation. We explain these properties as follows.

1.4.1.1 Adversarial Perceptibility

Generally, there are two types of perceptions for adversarial signals (Akhtar & Mian, 2018). In the first type, adversarial signals are perceivable by humans[34] (positive). However, the trained model cannot predict the correct output (negative). This type is commonly known as a false-negative adversarial example (Akhtar & Mian, 2018). In the second category (a.k.a. false-positive), the adversarial signal is neither perceivable by humans nor the recognition model.

[34] Either visually through observing weird patterns in the spectrogram or hearing strange noises upon playing the signal.

This thesis focuses on the latter type since it is more closely related to the real-life threats of adversarial attacks.

1.4.1.2 Model Accessibility

For cases where the adversary has access to the transcription model, architecture of the classifier, training dataset, tuning hyperparameters, etc., the developed attack is considered as a white-box algorithm (Athalye & Carlini, 2018; Gil, Chai, Gorodissky & Berant, 2019). The opposite scenario of this case constitutes a black-box attack (Ilyas, Engstrom, Athalye & Lin, 2018; Jiang, Ma, Chen, Bailey & Jiang, 2019).

Regarding cardinality, the total number of available publications on the white-box adversarial attack is relatively more than the black-box counterpart. This presumably has two incentives (Cheng, Dong, Pang, Su & Zhu, 2019):

1. developing a black-box adversarial attack is extremely more challenging than a white-box algorithm. Since in the black-box scenario the recognition model is not accessible, the adversary should:
    a. carefully approximate the weight vectors in order to yield a surrogate model (SM) to retrieve the decision boundary of the victim classifier (Uesato, O'donoghue, Kohli & Oord, 2018; Tang, Ma, Chen, Guo, Wang, Zeng & Zhan, 2020),
    b. reformulate the attack procedure toward the achieved SM.

    Both these two steps are challenging and impose considerable computational overhead to the entire adversarial optimization procedure.
2. The white-box attack development is mostly in regard to understanding the functionality of the data-driven models and their pitfalls. However, investigations on the black-box attacks address the complicated real-life applications (Goodfellow *et al.*, 2015).

In this thesis, we cover both white and black-box adversarial attacks developed against ESC and ASR models.

1.4.1.3 Output Specificity

The optimization formulation of the adversarial attack can be subject to a predefined output phrase (targeted) or towards any incorrect phrases other than the ground-truth (non-targeted). Technically, the computational cost[35] for the targeted attack is relatively greater than non-targeted (Poursaeed, Katsman, Gao & Belongie, 2018). However, targeted attacks are more in line with the real-life challenges.

1.4.1.4 Perturbation Measurement

Regarding the constraints mentioned in Section 1.4 about developing an adversarial attack, the crafted adversarial signal (or its representation) should be very similar to its associated ground-truth. Thus, for properly measuring this similarity, we need to employ an accurate metric as follows.

1. *Similarity metrics for 1D signals:* variants of dB-scale logarithmic similarity metric (Carlini & Wagner, 2018).
2. *Similarity metrics for 2D representations:* standard Lebesgue norms such as l_0, l_2, l_∞ (Szegedy *et al.*, 2014; Papernot, McDaniel, Jha, Fredrikson, Celik & Swami, 2016d; Carlini & Wagner, 2017b).

It has been demonstrated that the attack success rate is partially dependent on the choice of the similarity metrics mentioned above (Szegedy *et al.*, 2014). Hence, for effective adversarial attacking, an optimal metric should balance signal quality and attack fooling rate.

The following section provides general formulations for developing adversarial attacks against data-driven audio and speech recognition models.

---

[35] Mostly in terms of gradient computation (Masure, Dumas & Prouff, 2019) and the total number of required callback to the reference signal. For more details see Section 4.4.3.

1.4.2 Adversarial Attacks in Practice

Herein, we explain the formulation of adversarial attacks for both end-to-end and representation-level ESC and ASR systems.

1.4.2.1 Attacks for End-to-End Transcription Systems

In general, an end-to-end adversarial attack runs an optimization algorithm for $\langle \vec{x}_{\text{orig}}, \hat{\mathbf{y}}_i \rangle$ where $\vec{x}_{\text{orig}}$ stands for the original signal, and $\hat{\mathbf{y}}_i$ denotes the associated predefined target phrase (Carlini & Wagner, 2018):

$$\min_{\delta} \|\delta\|_F + \sum_i c_i L_i(\vec{x}_{\text{adv}}, \hat{\mathbf{y}}_i) \quad \text{s.t.} \quad l_{\text{dB}}(\vec{x}_{\text{adv}}) < \epsilon \tag{1.8}$$

and:

$$l_{\text{dB}}(\vec{x}_{\text{adv}}) = l_{\text{dB}}(\delta) - l_{\text{dB}}(\vec{x}_{\text{orig}}) \mid \vec{x}_{\text{orig}}, \vec{x}_{\text{adv}} \in \mathbb{R} \tag{1.9}$$

where $\vec{x}_{\text{adv}} = \vec{x}_{\text{orig}} + \delta$. Additionally, δ indicates the adversarial perturbation achievable through this iterative optimization formulation. Moreover, c_i is the hyperparameter for scaling the loss function $L_i(\cdot)$ regarding the length of the ground truth phrase $\mathbf{y}_i$ ($\mathbf{y}_i \neq \hat{\mathbf{y}}_i$). Furthermore, $l_{\text{dB}}(\cdot)$ computes the relative loudness of the signal in the logarithmic dB-scale, and ϵ is the audible threshold defined by the adversary.

Two typical pseudocodes (inspired from (Goodfellow *et al.*, 2015) and (Yakura & Sakuma, 2018)) for attacking end-to-end transcription systems are shown in Algorithm 1.1 and 1.2. There are several variants for this algorithm, and Eq. 1.8, where they often employ different loss functions, loudness metrics, adaptive scaling, etc. It has been shown that all these attack algorithms can debase the performance of the data-driven recognition models to almost zero (Carlini & Wagner, 2018).

Algorithm 1.1 A typical pseudocode for adversarial attack in the end-to-end framework (taken from Carlini & Wagner (2018)). Herein, $\mathcal{L}(\cdot)$ is the same as $L(\cdot)$.

Algorithm: Adversarial attack against end-to-end transcription systems.
Input: $\vec{x}_{\text{org}}$, $\mathbf{y}$, $\hat{\mathbf{y}}$
Output: $\vec{x}_{\text{adv}}$

$\vec{x}_c \leftarrow \vec{x}_{\text{org}}$; /* candidate adversarial signal */
while $\hat{\mathbf{y}} = \mathbf{y}$ **do**
    $\delta \leftarrow \min_\delta \|\delta\|_2 + \sum_i c_i \mathcal{L}_i(\vec{x}_c, \hat{\mathbf{y}}_i)$
    $\vec{x}_c \leftarrow \vec{x}_c + \delta$
end while
$\vec{x}_{\text{adv}} \leftarrow \vec{x}_c$; /* crafted adversarial signal */

Algorithm 1.2 Another typical pseudocode for adversarial attack in the end-to-end framework (taken from Yakura & Sakuma (2018))

Algorithm: Adversarial attack against end-to-end transcription systems.
Input: $\vec{x}_{\text{org}}$, $\mathbf{y}$, $\hat{\mathbf{y}}$
Output: $\vec{x}_{\text{adv}}$

$\vec{x}_c \leftarrow \vec{x}_{\text{org}}$; /* candidate adversarial signal */
while $\hat{\mathbf{y}} = \mathbf{y}$ **do**
    $\delta \leftarrow \min_\delta \|\delta\|_2 + \sum_i c_i \mathcal{L}_i(\vec{x}_c, \hat{\mathbf{y}}_i)$
    while $\delta > \iota_{ht}$ **do**
        Project δ using some transformations.
    end while
    $\vec{x}_c \leftarrow \vec{x}_c + \delta$
end while
; /* ι_{ht}: hearing threshold. */
$\vec{x}_{\text{adv}} \leftarrow \vec{x}_c$; /* crafted adversarial signal */

1.4.2.2 Attacks for Representation-Level Transcription Models

Developing an adversarial attack for a representation-level audio and speech recognition model is almost the same as in the computer vision domain. For instance, it can be formulated as an optimization problem toward achieving a very small perturbation δ, as stated by Szegedy *et al.*

Algorithm 1.3 A typical pseudocode for adversarial attack in the representation-level framework (taken from Goodfellow *et al.* (2015))

Algorithm: Adversarial attack against representation-level recognition systems.
Input: $\mathbf{x}_{\text{org}}$, $\mathbf{y}$, $\hat{\mathbf{y}}$
Output: $\mathbf{x}_{\text{adv}}$

$\mathbf{x}_c \leftarrow \mathbf{x}_{\text{org}}$; `/* candidate adversarial signal */`
while $\hat{\mathbf{y}} = \mathbf{y}$ **do**
  $\mathbf{x}_c \leftarrow \mathbf{x}_c + \delta \nabla_{\mathbf{x}} J(\mathbf{x}_c, \hat{\mathbf{y}}_i)$
end while
$\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x}_c$; `/* crafted adversarial signal */`

(2014):

$$\min_{\delta} \quad f^*(\underbrace{\mathbf{x}_{\text{org}} + \delta}_{\mathbf{x}_{\text{adv}}}) \neq f^*(\mathbf{x}_{\text{org}}) \tag{1.10}$$

where $\mathbf{x}_{\text{org}}$ and f^* denote the original spectrogram and the post-activation function of the victim classifier, respectively. Although interpreting a signal representation is very difficult for human eyes, still δ should not be perceivable.

Algorithm 1.3 (Goodfellow *et al.*, 2015) illustrates a typical pseudocode for attacking the recognition models trained on spectrograms. There are many variants for this algorithm and also Eq. 1.10, which incorporates different optimization policies such as (Goodfellow *et al.*, 2015):

$$\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x}_{\text{org}} + \delta \frac{\nabla_{\mathbf{x}_{\text{org}}} J(\mathbf{x}_{\text{org}}, \hat{\mathbf{y}}_i)}{\left\| \nabla_{\mathbf{x}_{\text{org}}} J(\mathbf{x}_{\text{org}}, \hat{\mathbf{y}}_i) \right\|} \tag{1.11}$$

where $\mathbf{x}_{\text{adv}}$ and $J(\cdot)$ denote the adversarial spectrogram and the Jacobian matrix containing the gradient vectors of the victim model, respectively. In this thesis, we cover such attack variants, which have been benchmarked for signal representations.

Almost all the recognition algorithms mentioned in Sections 1.2 and 1.3 use a type of signal representation (e.g., MFCC). This obliges us to explain how an adversarial spectrogram can negatively affect the performance of such systems.

Assume we have an ESC or ASR system that employs a front-end classifier trained on legitimate spectrograms. We show how crafted adversarial spectrograms can pose security concerns for these systems:

1. *White-box scenario:* the adversary has full access to the entire system details, including dataset, classifier architecture, potential tuning parameters, required hyperparameters, and the complete weight vectors. Therefore, the adversary can easily feed adversarial spectrograms to the model and fool it toward any target phrase or label.
2. *Black-box scenario:* the adversary does not have access to the system details mentioned above. In fact, the adversary can only input a 1D signal to the system and receive a predicted label or phrase. In this scenario, the adversary can reconstruct an audio signal from an adversarial spectrogram (with or without a surrogate model) and feed it to the system. Since the model is trained on spectrograms, the system first converts the input audio into a spectrogram, which also recovers the adversarial perturbation (this reconstruction does not impose technical difficulty since spectrogram and 1D signal are dual[36]). This spectrogram can also fool the victim model towards any wrong phrases defined by the adversary (Koerich, Esmailpour, Abdoli, Britto & Koerich, 2020).

In this section, we briefly reviewed the characterization of adversarial attacks for end-to-end and representation-level recognition models. The following section explains a taxonomy of the proposed approaches for defending ESC and ASR models against such adversarial attacks.

1.5 Adversarial Defenses for ESC and ASR Systems

As illustrated in Figure 1.2, there are generally two categories of approaches for securing ESC and ASR models against a variety of adversarial attacks (Akhtar & Mian, 2018; Zhang, Zhang & Zhang, 2019a; Mustafa, Khan, Hayat, Goecke, Shen & Shao, 2019; Hu, Shang, Qin, Li, Wang & Wang, 2019a; Yuan, He, Zhu & Li, 2019; Cohen, Sapiro & Giryes, 2020):

1. *Proactive:* this category includes all the approaches in which the defense policy is a part of the classifier development. In other words, proactive defense algorithms do not employ

[36] There are plenty of straightforward approaches for reconstructing one from another.

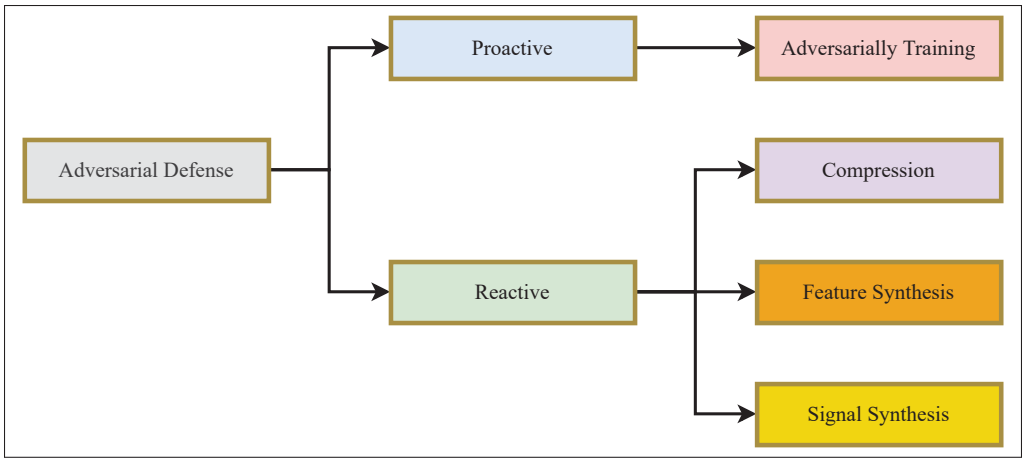

Figure 1.2 General taxonomy of adversarial defense for ESC and ASR systems

any postprocessing[37] operations for bypassing (in particular, fading the perturbation in the signal or the spectrogram) the potential adversarial perturbation. In fact, the recognition algorithm should implement some built-in strategies to yield a reliable model. One popular example of such proactive defenses is adversarially training[38] (Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand & Lempitsky, 2016; Tramèr, Kurakin, Papernot, Goodfellow, Boneh & McDaniel, 2017; Sun, Yeh, Hwang, Ostendorf & Xie, 2018).

2. *Reactive:* refers to the group of algorithms that incorporate an auxiliary model or a postprocessing module for potentially removing the adversarial perturbation. Synthesis-based defense algorithms[39] are among the popular reactive defense approaches since they do not obfuscate gradient information (Lee, Han & Lee, 2017; Samangouei *et al.*, 2018b; Bao, Liang & Wang, 2018; Athalye *et al.*, 2018b). We explain subcategories of reactive defenses in the following.

[37] Possibly preprocessing, as well.

[38] We succinctly explained this technique in the previous chapter. We also refer to Appendix II for more details.

[39] Additional explanations are available in Chapter 5 and Appendix III.

1.5.1 Reactive Adversarial Defense

Developing a proactive defense algorithm is fairly more challenging than a reactive approach. This is mainly due to integrating the defense policy into the classifier development procedure according to the definition of the proactive defense framework. This framework contributes to making a reasonable trade-off between recognition accuracy and model robustness. To the best of our knowledge, there is no investigation on proactive defense approaches in the context of ESC and ASR, and all the proposed defense algorithms fit in the reactive category[40]. This category includes three major subcategories as follows.

1.5.1.1 Compression

It has been shown that employing low-level transformation operations, to some extent, bypasses the adversarial perturbation on the signal (or the associated spectrogram). These operations include but are not limited to MPEG-layer-3 and multi-rate compressions (Pryadi, Gandi & Kanalebe, 2008; Ireland, Knuepffer & McBride, 2015; Das, Shanbhogue, Chen, Chen, Kounavis & Chau, 2018). However, according to our conducted experiments, these techniques might not be able to detect strong adversarial signals that have been carefully tuned during optimization.

1.5.1.2 Feature Synthesis

Synthesizing a feature vector similar to the representation of the given speech signal using an autoencoder-based GAN is another subcategory for the reactive defense approaches (Latif, Rana & Qadir, 2018). In fact, these generative models craft a new set of MFCC vectors for every given speech signal aiming at fading the potential adversarial perturbation during the synthesis (reconstruction) procedure. The major comment with the algorithms that fit in this defense subcategory is that they may not constitute reliable defense approaches since they transform the

[40] We introduce a proactive defense approach in Appendix II.

feature vectors using an autoencoder. Unfortunately, autoencoding exploits a filtration operation directly on the signal and it often results in gradient obfuscation[41] (Athalye *et al.*, 2018b).

1.5.1.3 Signal Synthesis

Generally, there are two subcategories for signal synthesis-based defense algorithms as the following:

1. *Explicit synthesis:* this subcategory is similar to the feature synthesis-based defense approaches since they run some filtration operations directly on the signal, however without an autoencoder (Song, Shu, Kushman & Ermon, 2018; Li, Zhang, Jia, Xu, Zhang, Wang, Ma & Gao, 2020a). According to Athalye *et al.* (2018b), these defenses might also give false senses of security against adversarial attacks.
2. *Implicit synthesis:* includes all the algorithms that do not run any filtration operation directly on the given input signal. In fact, they iteratively find a safe vector for the generative model to synthesize a signal (or its associated spectrogram) very similar to the given input speech (Samangouei *et al.*, 2018b). This group of algorithms meets all the defense reliability conditions discussed in the previous chapter.

Our focus in this thesis is on developing defense algorithms which fit in the latter subcategory. In the next section, we discuss some major challenges for developing reliable adversarial defense approaches for real-life applications.

1.5.2 Challenges

Although considerable progress has been made in adversarial defense over the last few years, there is still no fully functional approach for securing ESC and ASR systems against adversarial attacks. This poses a major security concern and obliges us to address the potential challenges and develop possible resolutions for such a situation. In the following, we briefly mention a few of those challenges.

[41] We refer to the previous chapter concerning the definition of the reliable defense approach.

1. To the best of our knowledge, there is no investigation on making a trade-off between recognition accuracy and robustness of the recognition model against adversarial attacks. This could have arisen from the inverse relation between them[42]. Therefore, we need to design better classification architectures and employ some signal transformation techniques to fill the gap between recognition accuracy and model robustness.
2. Developing an adversarial attack for audio and speech signals is computationally expensive. This is presumably because common attack optimization formulations do not necessarily yield a universal adversarial perturbation (Carlini & Wagner, 2018). Considering the high dimensionality of an input signal, perturbations should be optimized per frame (e.g., every 50 ms) and this adds considerable computational overhead to the attack formulation. Consequently, this makes the adversarially training procedure almost impossible for datasets including long-duration signals.
3. It has been demonstrated that playing an adversarial signal over the air might fade out the adversarial perturbation (Carlini & Wagner, 2018; Yakura & Sakuma, 2018). Hence, toward developing a fast attack algorithm for adversarially training purposes, its robustness over consecutive playbacks should be considered. Unfortunately, this aspect has been totally neglected in almost all the proposed attack and defense approaches.
4. According to our initial experiments, unfortunately, all the adversarial defense categories mentioned in Section 1.5 negatively affect the quality of the signals. Therefore, a reliable defense algorithm should keep a balance between recognition accuracy and model robustness and avoid degrading the quality of the input signals.
5. GAN is the standard generative model often used in reactive defense algorithms, and unfortunately, all of them, at different levels, suffer from instability and mode collapse issues (Brock, Donahue & Simonyan, 2019). Training a stable generative model improves the accuracy of the synthesis-based defense approaches. Thus, we need to devise a more functional regularization technique to enforce its generalizability.

[42] We address this relation in chapter 4.

In the following chapters, we address all the challenges mentioned above and propose approaches to tackle them properly.

1.6 Summary

In this chapter, we reviewed some important concepts about ESC and ASR systems. We firstly, explained three common representation techniques for reducing dimensionality and enhancing the quality of the signals, namely MFCC, STFT, and DWT. As stated, these spectral representations have become standard representations for the entire signal processing domains since decades ago. Secondly, we briefly reviewed the state-of-the-art algorithms developed for environmental sound classification, followed by a concise analysis of speech-to-text transcription systems. Following this, we characterized the existence of adversarial signals both in raw 1D and 2D representation levels. Finally, we explained the general taxonomy of adversarial defenses and highlighted some common challenges toward developing a reliable algorithm against a variety of attacks.

CHAPTER 2

UNSUPERVISED FEATURE LEARNING FOR ENVIRONMENTAL SOUND CLASSIFICATION USING WEIGHTED CYCLE-CONSISTENT GENERATIVE ADVERSARIAL NETWORK

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in « Elsevier Applied Soft Computing » in November 2019.

Abstract

In this paper we propose a novel environmental sound classification approach incorporating unsupervised feature learning via the spherical K-Means++ algorithm and a new architecture for high-level data augmentation. The audio signal is transformed into a 2D representation using a discrete wavelet transform (DWT). The DWT spectrograms are then augmented by a novel architecture for a cycle-consistent generative adversarial network. This high-level augmentation bootstraps generated spectrograms in both intra- and inter-class manners by translating structural features from sample to sample. A codebook is built by coding the DWT spectrograms with the speeded-up robust feature detector and the K-Means++ algorithm. The Random forest is the final learning algorithm which learns the environmental sound classification task from the code vectors. Experimental results in four benchmarking environmental sound datasets (ESC-10, ESC-50, UrbanSound8k, and DCASE-2017) have shown that the proposed classification approach outperforms most of the state-of-the-art classifiers, including convolutional neural networks such as AlexNet and GoogLeNet, improving the classification rate between 3.51% and 14.34%, depending on the dataset.

2.1 Introduction

Environmental sound classification has attracted the interest of several researchers in machine learning because of its vast applications (Chu *et al.*, 2009a; Radhakrishnan *et al.*, 2005; Xu, Xu, Duan, Jin & Luo, 2008). However, this is a challenging problem due to the complex nature of environmental sounds in terms of dimensionality, different mechanisms of sound production, overlapping of different sources, and lack of high-level structures usually observed in speech and in many types of musical sounds (Salamon & Bello, 2015b). This complex nature masked by natural acoustic noises (Noda, Mori, Ishibashi & Itomi, 1987) can make the classification of specific sounds very challenging. This challenge becomes more difficult when audio classes do not have similar sound production mechanisms such as "car horn" and "car engine idling" in open scenes like streets or parks (Chu *et al.*, 2009a; Ellis & Lee, 2004; Chaudhuri & Raj, 2013).

In the literature, the classification of environmental sounds has been addressed using both standalone and ensemble classification setups incorporating conventional classifiers and deep neural networks where the input signal can be represented by an audio waveform (1D) or converted to a mid-level representation (2D) such as a spectrogram (Salamon & Bello, 2015b; Aytar, Vondrick & Torralba, 2016; Dai, Dai, Qu, Li & Das, 2017; Mun, Park, Han & Ko, 2017; Piczak, 2015a; Salamon & Bello, 2017; Tokozume & Harada, 2017). The audio signal may also be represented by handcrafted features in the spectral or the cepstral domains mainly via frequency transformations which are lossy operations. Zero-crossing rate (Lu, Zhang & Jiang, 2002), spectral flux and centroid (Tzanetakis & Cook, 2002), chroma vector (Ellis, 2007), Mel frequency cepstral coefficients (MFCCs) (Logan et al., 2000), short-time Fourier transform (STFT) (Smith et al., 2011), cross recurrent plot (CRP) (Serra, Serra & Andrzejak, 2009), and discrete wavelet transform (DWT) (Van Fleet, 2011) are among the most well-known handcrafted features for audio classification (Papakostas, Spyrou, Giannakopoulos, Siantikos, Sgouropoulos, Mylonas & Makedon, 2017). These handcrafted features not only reduce the dimensionality of the audio signal but may also reduce some types of noise and help to extract time-varying descriptors which provide a better discrimination. The approaches that use these features have shown relatively better performance than the approaches that use 1D signals directly in both

classification and clustering tasks, mainly when employing conventional classifiers such as support vector machines (SVMs) (Gerek & Ece, 2008).

MFCC is a common and reliable informative representation format for analyzing audio and for this reason, most of the proposed classification approaches in this domain rely on it (Radhakrishnan *et al.*, 2005; Cai, Lu, Hanjalic, Zhang & Cai, 2006; Ganchev, Fakotakis & Kokkinakis, 2005; Heittola, Mesaros, Eronen & Virtanen, 2013). MFCCs are handcrafted features based on the human auditory system, which can make a reasonable balance between handling the complex nature of real-life sounds and providing informative feature vectors for classification purposes. In addition to traditional classifiers such as Gaussian mixture models (Godino-Llorente, Gomez-Vilda & Blanco-Velasco, 2006), hidden Markov models (Gales & Young, 1992) and K-nearest neighbor (Eronen, Peltonen, Tuomi, Klapuri, Fagerlund, Sorsa, Lorho & Huopaniemi, 2006), convolutional neural networks (ConvNets) (Deng, Hinton & Kingsbury, 2013) have been evaluated on MFCC feature vectors and achieved better results than the 1D audio signal. However, MFCCs have shown to be very sensitive to background noise and this might affect the performance of classifiers for noisy environmental sounds (Cotton & Ellis, 2011).

With the recent advances in deep learning, many strong classifiers such as ConvNets have been introduced, which are designed to learn directly both from 1D and 2D data. ConvNets are quite similar to dense deep neural network (DNN) where the main difference is the inclusion of convolution layer(s) to deal with raw data. The main advantage of these networks is their ability to learn directly from raw data rather than handcrafted features. ConvNets have been used with audio waveforms with several convolution layers incorporating different 1D signal augmentation methods (Salamon & Bello, 2017). Experimental results have shown competitive accuracy compared to unsupervised sound classifiers (Salamon & Bello, 2015b), ConvNets on MFCCs (Piczak, 2015b), and even better performance (Palaz, Doss & Collobert, 2015) depending on the dataset. ConvNets have also been evaluated with a combination of 1D and MFCC feature vectors which resulted in low classification error (Tokozume & Harada, 2017). This shows the importance of the representation space in extracting discriminating features.

Audio signals are high dimensional, which means that more than a thousand real values need to be used to represent a short audio signal. Due to this fact, it is preferred to train classifiers on 2D audio representations over audio waveforms. However, ConvNets have shown great classification performances in 1D signal format (Salamon & Bello, 2017; Abdoli, Cardinal & Koerich, 2019), so far, they could not outperform AlexNet and GoogLeNet on STFT, DWT, and CRP spectrograms (Weiping, Jiantao, Xiaotao, Xiangtao & Shaohu, 2017). The majority of recent papers in audio classification especially environmental sounds are on 2D representations mainly for DNNs such as the networks introduced in (Deng *et al.*, 2013). STFT, DWT, and CRP are the main approaches for producing spectrograms and they can also be combined to augment the amount of data and to extract more informative 2D representations for training ConvNets (Boddapati *et al.*, 2017). It has been shown that STFT and DWT have more competence for extracting temporal and structural content for ConvNets (Wyse, 2017). For some common environmental datasets, GoogLeNet and AlexNet have achieved the highest recognition accuracy with quite high confidence.

However, one of the main bottlenecks for using ConvNets in environmental sound classification is the amount of data required to train such networks properly due to the high number of parameters to adjust. The two main approaches that have been used to circumvent this problem are: (i) fine-tuning ConvNets pre-trained on other domains/datasets; (ii) generating artificial samples by data augmentation. Both 1D and 2D data augmentation approaches (Weiping *et al.*, 2017; Salamon & Bello, 2015a) have been proposed for improving classification performance which proves the importance of providing better input rather than implementing highly complex and costly networks (Mun *et al.*, 2017). There are several algorithms for augmenting a dataset both in terms of enhancing samples' visual quality and quantity. Augmentation in 2D representations like spectrograms is mostly being implemented with low-level transformations (Cireşan, Meier, Masci, Gambardella & Schmidhuber, 2011) including translation, shearing, rotation, scaling, aspect ratio, flipping, etc., which in general may not improve the performance of conventional or deep learning classifiers. The linear nature of these affine transformations may not cause a high impact on the classifier decision boundaries (Zhu *et al.*, 2018c). It is worth mentioning that even

these low-level data augmentations have been sometimes contributed significantly in training ConvNets and reducing overfitting.

Elastic deformation (Simard, Steinkraus & Platt, 2003) is another type of low-level augmentation which has been used in spectrograms. The elastic deformation implements a similarity transformation which interpolates between highly correlated spectrogram sub-manifolds. However, when the resolution of the spectrogram is small, and it does not have much active areas[43] (super uniform areas in pixel-wise level), this augmentation may not work well especially for deep learning models. Extracting covariant patches and color space channel intensity alteration (Krizhevsky *et al.*, 2012) as well as other types of pixel-level augmentation scheme has been utilized in many spectrogram classification tasks. In addition to the linear nature of low-level augmentation, they cannot enhance data distribution which is usually determined by high-level features. Some methodologies have been proposed for circumventing this problem such as learning multivariate normal distribution for each class with respect to their mean manifolds (Hauberg, Freifeld, Larsen, Fisher & Hansen, 2016). Implementation of this augmentation in the real world, especially for long audio sequences of high dimension is not optimal. One potential solution could be multivariate distribution learning in representation space (Dixit, Kwitt, Niethammer & Vasconcelos, 2017) with respect to the structural components (Wang, Bovik, Sheikh & Simoncelli, 2004) of a spectrogram.

In this paper, we propose a novel architecture for data augmentation which translates one spectrogram to another using a generative model named Weighted Cycle-Consistent Generative Adversarial Network (WCCGAN), as well as a novel approach for environmental sound classification based on unsupervised feature learning. The proposed approach has four main steps: (i) audio dimension conversion and preprocessing (from 1D to 2D); (ii) data augmentation using the proposed WCCGAN; (iii) extracting feature vectors from the augmented dataset via speeded-up robust feature detector (SURF) algorithm and learning a codebook of representative codewords; and (iv) training a random forest algorithm on code vectors. The experimental results have shown that our approach outperforms cutting-edge classifiers such as AlexNet and

[43] Areas with uniform distribution and low intensities.

GoogLeNet in four benchmarking environmental sound datasets: ESC-10 (Piczak, 2015b), ESC-50 (Piczak, 2015b), UrbanSound8k (Salamon *et al.*, 2014a) and DCASE-2017 (Mesaros, Heittola, Diment, Elizalde, Shah, Vincent, Raj & Virtanen, 2017). Besides that the experimental results have also shown the remarkable performance of the proposed data augmentation approach for both the unsupervised feature learning and supervised approaches.

The organization of this paper is as follows. In Section 2.2 we discuss the transformation of audio waveforms (1D) into spectrograms (2D), as well as the preprocessing steps preceding and succeeding such a transformation for data augmentation purposes. Section 2.3 presents the WCCGAN for high-level spectrogram augmentation. In Section 2.4, we explain our feature learning methodology using SURF descriptors and the spherical K-Means++ algorithm, and also the classification approach based on random forests. Section 2.5 provides details about the architecture of the proposed WCCGAN and the experiments carried out in four benchmarking datasets. In Section 2.6 we compare the importance of pitch-shifting as one of the basic data augmentation approaches over all other algorithms presented in (Salamon & Bello, 2017). We show that implementing all types of data augmentations does not necessarily produce informative features favorable to the proposed classifier. We also compare the performance of the proposed WCCGAN with the cycle-consistent GAN proposed by Zhu et al. (Zhu *et al.*, 2018c) to emphasize the importance of adapting generative model architectures according to the application. The conclusions and future work are presented in the last section.

2.2 Preprocessing and Spectrogram Generation

In this section we present the preprocessing steps to artificially expand the size of an audio dataset (i.e., augmentation) by creating modified versions of the audio clips and the strategy used to convert such audio clips into spectrograms.

2.2.1 1D Data Augmentation

Given the relatively small size of the environmental sound datasets, one of the recommended steps before transforming an audio waveform to a 2D representation is to boost the amount, distribution, and cardinality of the samples of each class in the datasets. Data augmentation can be carried out by applying some filters on an audio signal such as pitch shifting, time stretching, compressing dynamic range, and background noise removal (Salamon & Bello, 2017). These operations can be individually applied to an audio sample to produce slightly modified versions of it and increase the number of samples. Finally, these crafted samples are added to the original dataset. Since these augmentation filters increase the number of samples of the dataset, they may have potential to affect the performance of data-driven classifiers.

It has been shown (Boddapati *et al.*, 2017) that the pitch-shifting filter alone can highly boost the quality of audio recordings when compared with applying all above-mentioned augmentation filters as proposed in (Salamon & Bello, 2017). After conducting several exploratory experiments, we have found out that for most of the environmental sound datasets applying all 1D data augmentation filters do not necessarily produce good audio samples in terms of producing samples with low inter-class and high intra-class similarity. In Section 2.6 we show some experimental results that support this claim. Therefore, we only use the pitch-shifting augmentation as its constructive effects have been shown in (Salamon & Bello, 2017) for both supervised and unsupervised feature learning. For such an aim, we use static pitch shifting scales (McFee, Humphrey & Bello, 2015a). This boosts the number of samples in the dataset with respect to the number of applied scales.

2.2.2 Spectrogram Generation

STFT, DWT, and CRP are the main approaches for producing spectrograms for an audio signal. ConvNets have shown strong capability in learning from these spectrograms either standalone (Weiping *et al.*, 2017) or pooled together (Boddapati *et al.*, 2017). The DWT representation is more stable to time warping deformations and it can better characterize time

varying structures compared to other representations such as STFT (Mallat, 2012). The STFT transformation is somewhat similar to DWT in terms of producing low and high frequency components encoded as spectrograms. Considering a discrete-time signal $x[n]$, its DWT transformation is given by:

$$\mathrm{DWT}(x[n]) = (x[n] \otimes g[n]) = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \tag{2.1}$$

where $\otimes$ denotes the convolution of $x[n]$ and $g[\cdot]$ (mother function which produces other signals which can be either low or high pass filter sets). This operation can be applied to at most the minimum length of $x[n]$. The 2D representation of this signal can be computed by:

$$\mathrm{S_{DWT}}\{x[n]\} \equiv |\mathrm{DWT}(x[n])|^2 \tag{2.2}$$

where S_{DWT} denotes the spectrogram of the signal $x[n]$.

We generate DWT spectrograms using our modified version of the sound explorer C++ script (Hanov, 2008) for the original audio signals as well as the pitch-shifted audio samples to handle audio clips with any length (time duration).

2.2.3 Spectrogram Enhancement

Each generated spectrogram is a 2D array of intensity values which can be noisy when its associated audio signal is affected by environmental noise(s). In this case, adjusting the distribution of the intensity values can help to extract/learn more informative features (Segura, Benítez, Torre & Rubio, 2002). For improving the color space and the dynamic color contrast of the intensity values (as part of data augmentation pipeline (Park, Chan, Zhang, Chiu, Zoph, Cubuk & Le, 2019)), we apply a histogram equalization filter (Gonzalez & Woods, 2002). Considering each pixel intensity of the generated spectrogram S as $S(i, j)$, then the enhanced spectrogram (S_{heq}) is defined in Equation 2.3.

$$S_{heq}(i,j) = \left\lfloor (s-1) \sum_{i=0}^{S(i,j)} p_i \right\rfloor \tag{2.3}$$

where s is the supremum of 8-bit precision and p_i denotes the ratio of pixels with intensity i over the total number of pixels. This filter expands the intensity range of a given spectrogram in a balanced distribution. In the next section we explain how to structurally augment the generated spectrograms towards more informative samples.

2.3 Weighted Cycle-Consistent Generative Adversarial Network (WCCGAN)

The approach proposed for augmenting generated spectrograms is based on the Cycle-Consistent Generative Adversarial Network (CCGAN) which maps one spectrogram to another spectrogram. The efficiency of this 2D-to-2D translation has been proven in the literature for image-to-image translation problems (Isola, Zhu, Zhou & Efros, 2017; Zhu, Park, Isola & Efros, 2017a). The proposed GAN architecture is inspired by Zhu et al. (Zhu *et al.*, 2017b) with two main differences: (i) it incorporates two identity mapping functions for avoiding the oversmoothing of generated spectrograms, which affects the performance of the discriminator towards a wrong label other than the pre-defined target label; (ii) it employs different architectures for both generator and discriminator.

The proposed augmentation pipeline is implemented only in 2D space since mapping 1D-to-1D audio signals for augmentation purposes is very challenging due to the high dimensionality of audio signals. Our perspective in data augmentation is directed towards increasing inter- and intra-class structural contents over low-level pixel augmentations. This can help classifiers to reach a finer decision boundary among data sub-manifolds with minimum overlap. A more accurate way to impose structural contents on data augmentation is by using GAN since we can consistently control the mapping process from one image to another by adding an extra constraint to its loss function(s).

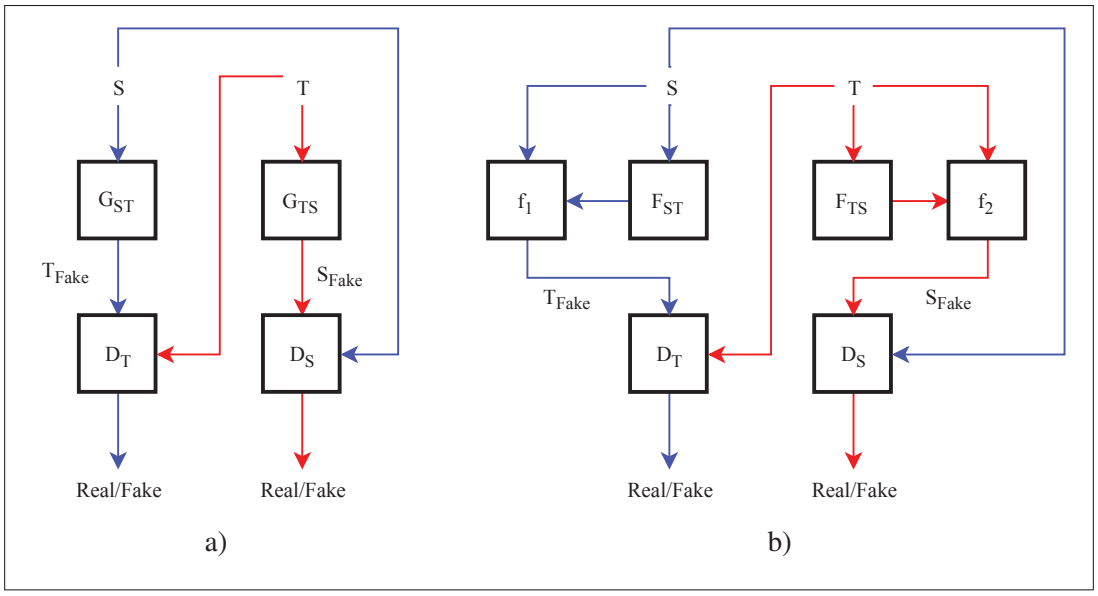

Figure 2.1 (a): Illustration of the original Cycle-Consistent GAN (CCGAN) for image-to-image translation where the cycle consistency imposes $G_{ST}(S_{Fake}) \approx T$ and $G_{TS}(T_{Fake}) \approx S$. (b): The proposed Weighted Cycle-Consistent GAN (WCCGAN) inspired by Zhu et al. (Zhu *et al.*, 2017b). Generators in our framework are F_{ST} and F_{TS} equivalent to G_{ST} and G_{TS}, respectively.

The original architecture of the CCGAN (Zhu *et al.*, 2017b) is shown in Figure 2.1(a) and it consists of two networks, one generator (G) and one discriminator (D) that capture data distribution and estimate the probability that a sample comes from the training data rather than G, respectively. G in a standard GAN generates fake data from latent variables with respect to the distribution of real training data, whereas in CCGAN (Zhu *et al.*, 2017a) it bijectively translates an input sample from a source S to a target T. In other words, this type of GAN has two generators and two discriminators which are trained independently.

In this paper we focus on both paired (when S is similar to T; or equivalently intra-class translation) and unpaired (when S and T are somewhat similar to each other, or equivalently inter-class translation) CCGAN. For the latter, we propose a pipeline for properly selecting source and target spectrograms with respect to the confusion matrix of the classifier. This high-level augmentation transfers structural components from the source spectrogram S to the

target spectrogram T. If the CCGAN is trained carefully, it can produce spectrogram samples that may help improve the performance of a classifier trained with such samples.

Producing realistic (natural-looking) spectrograms is not one of our priorities since any sort of spectrogram does not have much meaning for human eyes. Interestingly, spectrograms generated using our generator network may not produce samples similar to a given source, but discriminator shows reasonable sensitivity to it (matches the target label). Some examples of the generated samples are depicted in Figure 2.4. Forcing generators to produce very similar samples will result in divergences in the cycle consistency optimization. This condition for CCGANs mostly applies for augmenting datasets to which the human eyes perceive some structure like MNIST and ImageNet datasets.

In Figure 2.1(a), G_{ST} and G_{TS} stand for generators translating samples from $S \rightarrow T$ and $T \rightarrow S$, respectively. D_T and D_S denote the modules for discriminating real samples from generated fake samples from G_{ST} and G_{TS}. This can be achieved by optimizing the following criterion:

$$G_{S \rightarrow T} = \arg \min_{G_{S \rightarrow T}} \max_{D_T} \mathcal{L}_{GAN}(G_{S \rightarrow T}, D_T) \tag{2.4}$$

where the loss function $\mathcal{L}_{GAN}$ is defined in Equation 2.5.

$$\mathcal{L}_{GAN}(G_{S \rightarrow T}, D_T) = \mathbb{E}_{t \sim p_{target(t)}} \left[\log D_{T(t)}\right] + \mathbb{E}_{s \sim p_{source(s)}} \left[\log(1 - D_{T(t)}(G_{S \rightarrow T(s)}))\right] \tag{2.5}$$

where $p_{target(t)}$ and $p_{source(s)}$ denote the sample distributions in the target T and source S, respectively. The common problem with this definition of loss function is gradient vanishing which makes training and convergence almost impossible (Arjovsky, Chintala & Bottou, 2017). To circumvent this problem, in the proposed WCCGAN architecture depicted in Figure 2.1(b), we use a least-square loss function for GAN ($LSGAN$) as proposed in (Mao, Li, Xie, Lau, Wang & Smolley, 2017) for different domains S and T as given in Equations 2.6 and 2.7:

$$\mathcal{L}_{LSGAN}(F_{S \rightarrow T}, D_T) = \mathbb{E}_{t \sim p_{target(t)}} \left[(D_{T(t)} - 1)^2\right] + \mathbb{E}_{s \sim p_{source(s)}} \left[D_{T(t)}(F_{S \rightarrow T(s)})^2\right] \tag{2.6}$$

$$\mathcal{L}_{LSGAN}(F_{T\to S}, D_S) = \mathbb{E}_{s\sim p_{target(s)}}\left[(D_{S(t)} - 1)^2\right] + \mathbb{E}_{t\sim p_{source(t)}}\left[D_{S(s)}(F_{T\to S(t)})^2\right] \tag{2.7}$$

Though these loss functions minimize the approximated Jensen-Shannon divergence between two distributions of legitimate and generated data (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville & Bengio, 2014), they oversmooth the spectrograms. Oversmoothing affects the performance of the discriminator towards a wrong label other than the pre-defined target label. For rectifying this problem, we bypass the inputs to the discriminator. Hence, we add the two modules (f_1 and f_2) as depicted in Figure 2.1(b), which act as weighted bypasses (identity mapping) to the discriminators. The definitions of these two modules are provided in Equations 2.8 and 2.9.

$$f_1 = c_1 \odot S + \mu \odot F_{ST} \tag{2.8}$$

$$f_2 = c_2 \odot T + \sigma \odot F_{TS} \tag{2.9}$$

where dimensions of the generators and the input/target are bilinearly interpolated to match each other. The $\odot$ denotes the element-wise multiplication. The values of the constants c_1 and c_2, and the variables μ and σ are obtained empirically upon several experiments. Basically, f_1 and f_2 bypass connections have two main advantages in the proposed high-level augmentation setup. First, the low-to-high compensations because the regular CCGAN (Figure 2.1(a)) translates a randomly picked distribution from a low-dimension (e.g., pixel-level noisy sample) to a higher dimension which is a realistic image. Assuming that the dimension of the random drawn distribution is not very large, and no optimization overhead is involved (in the case of an optimal generator (Hoang, Nguyen, Le & Phung, 2018)), then potentially the following cycle-consistency criterion can yield a realistic fake sample:

$$\begin{aligned}\mathcal{L}_{cycle}(G_{S\to T}, G_{T\to S}) = \mathbb{E}_{s\sim p_{source(s)}}\left[\|G_{T\to S}(G_{S\to T}(s) - s)\|_1\right] + \\ \mathbb{E}_{t\sim p_{target(t)}}\left[\|G_{S\to T}(G_{T\to S}(t) - t)\|_1\right]\end{aligned} \tag{2.10}$$

where $\|\cdot\|_1$ is the L_1 norm. This might converge to a saddle point (where the minimax game in GAN is over) when the Kullback-Leibler divergence $\mathbb{KL}(p_{source(s)}, p_{target(t)}) \approx \mathbb{KL}(p_{target}, p_{source})$. In other words, the similarity between the source and the target distribution

should be high. When the similarity between samples is not high enough especially when they have been drawn from different classes, Equation 2.10 can no longer result in realistic fake images. f_i bypasses can overcome this problem by providing more information from a given legitimate input.

The second advantage of embedding f_1 and f_2 into the proposed WCCGAN is the ability of sharpening features that may have been oversmoothed during translation (especially in the discriminator domain). Finally, the total loss criterion which is optimized in our augmentation scenario is given in Equation 2.11:

$$\begin{aligned} \mathcal{L}_{total}(F_{S\to T}, F_{T\to S}, D_S, D_T) = \mathcal{L}_{LSGAN}(F_{S\to T}, D_T) + \\ \mathcal{L}_{LSGAN}(F_{T\to S}, D_S) + \alpha \mathcal{L}_{cycle}(F_{S\to T}, F_{T\to S}) \end{aligned} \tag{2.11}$$

where α is a scaling parameter for balancing the cycle whose value is also set manually upon experiments.

2.3.1 ConvNet Architecture for the Weighted Cycle-Consistent GAN

Assuming that we generate DWT spectrograms of 768×384 pixels, for high-level augmentation using the WCCGAN, we propose the architectures illustrated in Figure 2.2 for the generators ($F_{S\to T}$, $F_{T\to S}$). We started with a complex ConvNet model based on the AlexNet architecture for all four networks (generators and discriminators) of Figure 2.1(b), and we simplified this network by removing some layers which resulted in a simpler ConvNet architecture with 30% fewer parameters than AlexNet. Furthermore, when the architectures of discriminators and generators are similar, the cycle-consistency loss function follows a smooth and convex descending track. Therefore, we proposed two equivalent discriminators for both source-to-target and target-to-source mappings. The residual network shown in Figure 2.2 may have from three to seven residual blocks (He, Zhang, Ren & Sun, 2016), depending on the dataset. Each residual block contains two convolution layers and one bypassing residual connection. In all the layers depicted in Figure 2.2 the convolution layers have receptive field of 3×3 and stride 1×1. Also, the sizes of the generated outputs in the residual blocks are bilinearly interpolated to match each

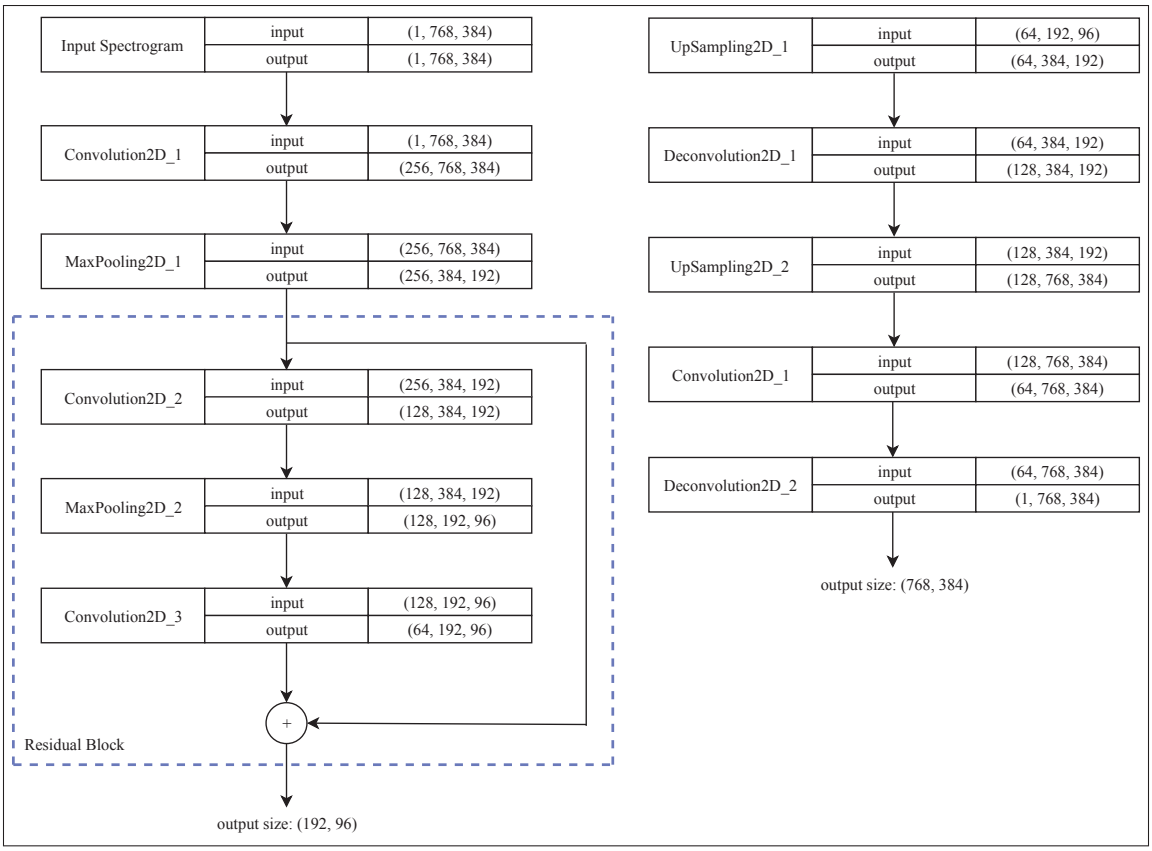

Figure 2.2 Generator architectures for DWT spectrograms: **left:** $F_{S\to T}$, and **right:** $F_{T\to S}$. Values inside of parentheses indicate the number of filters, height, and width of the spectrogram, respectively.

other. These architectures are not general and they need to be adapted depending on the type of problem and dataset. For discriminator functions D_T and D_S we use a single architecture as depicted in Figure 2.3. In the proposed architecture we also have receptive field of 3×3 and strides 1×1 and 2×2 for the first and second convolution layers, respectively. There is no generic way for determining an optimal structure for these two networks and we have basically relied on our initial experiments on the UrbanSound8k dataset. Changing the structures of these networks might affect the performance of image-to-image translation and it probably needs additional modifications/tuning of the hyperparameters. Therefore, we used the same architecture for the other datasets, but we have optimized the hyperparameters. Even if such an architecture is not customized to the other datasets, we have achieved good results as we show in Section 2.5.

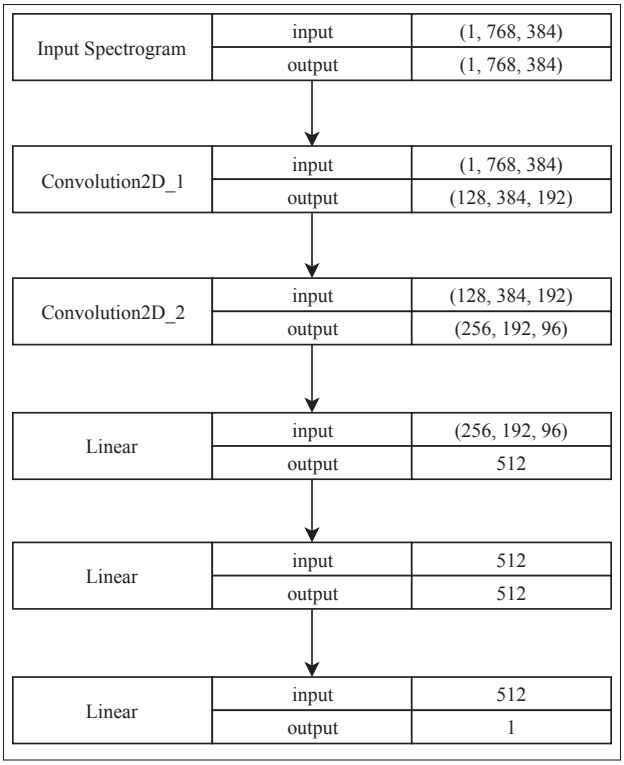

Figure 2.3 Network architecture for D_T and D_S

2.4 Unsupervised Feature Learning and Classification

The proposed approach for classifying DWT spectrograms is based on an unsupervised feature learning approach. The motivation behind proposing a shallow approach instead of a deep architecture as a front-end classifier is twofold. First, it has been shown that advanced deep neural networks such as AlexNet, GoogLeNet and other recent architectures are highly vulnerable to adversarial attacks as they can predict wrong labels with high confidence (Esmaeilpour *et al.*, 2020; Szegedy *et al.*, 2014). Secondly, conventional classifiers such as SVMs and RFs, which learn from handcrafted features are considerably more robust against such adversarial attacks than deep learning models (Esmaeilpour *et al.*, 2020). Taking advantage of these two facts, we propose a conventional data-driven model as front-end classifier and use a generative model based on a deep architecture as a back-end classifier for data augmentation purposes only. Therefore, the deep architecture helps the front-end classifier to learn more discriminant

boundaries. In this section, we present how to extract features from spectrograms and learn a codebook of representative codewords.

2.4.1 Feature Encoding

For extracting feature from the spectrograms, the speeded up robust feature (SURF (Bay, Tuytelaars & Van Gool, 2006)) is implemented, which is the modified version of the scale invariant feature transform (SIFT) (Lowe, 1999) by fast approximation of a Hessian matrix (for encoding principal curvatures at each point of interest) and producing integral images from spectrograms. Upon several experiments, SURF visual words from DWT provide us with better feature vectors compared to MFCC visual words which have been studied for music and environmental sound classification (Salamon & Bello, 2015b; Vaizman, McFee & Lanckriet, 2014). In Section 2.5 we provide some additional result in extracting SURF features from MFCC.

Each integral image represents the summation of the spectrogram pixels of a rectangular region with different sizes to produce local features. Using the box filter (for Gaussian approximation), SURF approximates the location and scale of each point of interest by using the determinant of the weighted Hessian matrix as the following.

$$H(\mathrm{p}, \sigma) \approx \begin{bmatrix} \hat{L}_{xx}(p, \sigma) & \hat{L}_{xy}(p, \sigma) \\ \hat{L}_{xy}(p, \sigma) & \hat{L}_{yy}(p, \sigma) \end{bmatrix} \tag{2.12}$$

$$\det(H(\mathrm{p}, \sigma)) = \hat{L}_{xx}(p, \sigma)\hat{L}_{yy}(p, \sigma) - [0.9(\hat{L}_{xy}(p, \sigma))]^2 \tag{2.13}$$

where $\hat{L}_{..}(p, \sigma)$ is the convolution of the second derivative of Gaussian with the spectrogram $S(x, y)$ at point x, and σ is the Gaussian scale (scale at which the point has been detected). After locating the interest points in space and scale, the SURF descriptor can be generated.

Assuming once again that we generate DWT spectrograms of 768×384 pixels, we divide each spectrogram into 16 sub-regions (4×4 grids of size 4×4) and compute Haar wavelet responses for obtaining orientation of interest points. In each sub-region, we compute a four-element

descriptor vector as given by Equation 2.14:

$$\texttt{descriptor}_{subregion} = \left[\sum dx, \sum dy, \sum |dx|, \sum |dy|\right] \tag{2.14}$$

The length of the regional feature descriptor is 16×4 which is represented by a 64-dimensional vector. These values are determined empirically on the UrbanSound8k dataset and they do not change during the implementation or across datasets. The majority of the settings for feature extraction are the default parameters of the OpenCV Library. For detecting interest points, a blob detector based on the Hessian matrix is implemented. Different Hessian threshold values have been evaluated, ranging from 250 to 1 000 on 15% of randomly selected samples of the dataset with four trials and 400 was set as a fair average threshold with respect to the performance of our classifier. Roughly, about 900 keypoints have been detected in each spectrogram. We have employed a non-maximum suppression strategy with a threshold of 0.6 to rectify the problem of detecting too many features. We skipped subregions in which SURF could not detect any feature. More details are presented in Section 2.5.

High resolution spectrograms to some extent can help SURF to extract more meaningful features but does not necessarily increase the performance in classification. Our main emphasis in this paper is the high-level augmentation which basically maps one sample to another aiming at increasing intra-class similarity and inter-class dissimilarity, regardless of the quality of the spectrogram. Resizing resolution of spectrograms, which perhaps changes the size of sub-regions, slightly affects the quality of the extracted features. For spectrograms of higher resolution (for instance 1152×576), we suggest increasing the dimension of feature vectors to 128 as our initial experiments have shown its positive impact on the final classification performance. In the next step, we learn a codebook of representative codewords (a.k.a visual words) from such feature vectors.

2.4.2 Organizing Visual Words into a Codebook Using Spherical K-Means++

The number of feature vectors extracted from the spectrograms is tremendously high and this negatively affects classifier's performance. Therefore, representing these vectors with respect to their similarities into centers and organizing them into a codebook can considerably improve the classification process.

We use the K-Means++ algorithm (Arthur & Vassilvitskii, 2007) as an unsupervised feature learning for organizing codewords. This clustering algorithm is adapted from the traditional K-Means algorithm where K denotes the number of potential seeds (centroids). This value is usually larger than the dimensionality of the audio data. The main advantage of the K-Means++ algorithm over the traditional K-Means algorithm is that it uses a weighted probability distribution over the data point (feature vector in our case) sub-manifold(s) with probability proportional to its squared distance to its neighbors. This is very useful in our case since feature vectors are not extracted from solid images. Similar to the traditional K-Means, the K-Means++ algorithm has a super polynomial structure and it might result in null seeds for similar data points (Dhillon & Modha, 2001). One possible solution provided for K-Means is adding an extra optimization constraint by binding seeds to have a unit L_2 norm which forces the centroid to roll over a unique sphere. This algorithm is called spherical K-Means (Coates & Ng, 2012). By taking advantage of this extra constraint and embedding it into the K-Means++ clustering algorithm, spherical K-Means++ (Endo & Miyamoto, 2015) turns out. The performance of standard spherical K-Means is studied for specific forms of environmental sound datasets with quite small cardinality (Stowell & Plumbley, 2014). It has been proven that this clustering algorithm produces competitive results with cutting-edge clustering and other advanced supervised classifiers (Dieleman & Schrauwen, 2013). Adding a spherical constraint in the distance objective function of K-Means usually results in improving the consistency in producing centroids.

Considering the feature vectors of an input spectrogram represented as a $X_{m,n}$ matrix where m and n denote the number of feature vectors and their dimensionality in the form of 1×n,

respectively ($1 \leq i \leq m$ and $1 \leq j \leq n$). In Equation 2.15 we define z_i for storing the assigned value (mean of centroids) of our K clusters which forms the matrix Z. Finally, our codebook is defined as $V \in \mathbb{R}^{n \times K}$.

$$z_j^i := \begin{cases} V^j x^{i^\top} & \text{if } j = \arg\max_l \left| V^l x^{i^\top} \right|_{j,i} \text{ and } \quad p(x) \sim \mathcal{N}(0, d^2(x^i)) \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

where x^i is a row from X and c is a constant value for weighting the square distance of each x^i to its nearest center. Specifically, d and p denote the distance between two feature vectors and their joint probability distribution, respectively, and $\top$ indicates matrix transposition. More details about the basics of the spherical K-Means algorithm is provided in (Coates & Ng, 2012). Finally, the two operations of Equation 2.16 update the centroids and normalize them by the L_2 norm, respectively. The centroids can be randomly normalized following a normal distribution. The codebook matrix V contains K organized clusters that we use to encode the training data and train a classifier.

$$V := XZ^\top + V, \qquad V^j := \frac{V^j}{\|V^j\|_2} \quad \forall j \tag{2.16}$$

2.4.3 Classification

For classifying the code vectors encoded against the codebook, we have considered the most performing conventional approaches, namely SVM with different kernels (linear, polynomial, radial basis functions) and random forest (RF). We decided to use a RF as our front-end classifier based on the recognition accuracy. We use the random forest (RF) algorithm (Breiman, 2001) with a different number of trees. This algorithm is an estimator which fits some decision trees on different sub-samples of given code vectors via averaging. We train this algorithm with different sizes of trees (estimators) with respect to the dimensions of the generated code vectors. For splitting a random tree node, the Gini impurity criterion is used as follows:

$$G = \sum_{i=1}^{n} p_i(1 - p_i) \tag{2.17}$$

where n denotes the number of classes in the target variable and p_i is the ratio of picking a random sample from class i. The maximum depth of trees varies from 16 to 64 with respect to the type of codebook. Specifically, for code vectors associated with long audio recordings, we use deeper trees. The minimum number of samples required to split an internal node is set to $0.02 \times m$ where m stands for the number of samples per class. This classifier has shown great potential for classifying code vectors (Salamon, Jacoby & Bello, 2014b). Upon our initial experiments, we have noticed that spherical binding of code vectors for the K-Means++ outperforms the standard one.

2.5 Experimental Results

We assess the performance of the proposed approach in four environmental sound datasets: UrbanSound8k, ESC-10, ESC-50, and DCASE-2017. The first dataset includes 8 732 audio samples of up to four seconds in duration distributed in 10 classes: air conditioner (AI), car horn (CA), children playing (CH), dog bark (DO), drilling (DR), engine idling (EN), gunshot (GU), jackhammer (JA), siren (SI), and street music (SM). The ESC-50 includes 2 000 samples of 5-second duration distributed in 50 classes including major groups of animals, natural sound clips and water sounds, human non-speech sounds, domestic sounds, and exterior noises. The ESC-10 is a subset of ESC-50 which includes 400 excerpts arranged in 10 classes: dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, fire crackling. Finally, DCASE-2017 consists of 4 680 10-second audio samples from 15 classes: bus, cafe, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office (multiple persons), residential area, train, tram, and urban park. Though the cardinality of samples per class in UrbanSound8k is not balanced, such a dataset contains the most challenging environmental sounds in real life compared to the other three datasets in terms of including different sound production mechanisms.

For low-level data augmentation, there is no automatic approach for tuning the pitch-shifting hyperparameter (t) and this depends on the type of audio signal, as mentioned in Section 2.1. Therefore, we have carried out a grid search starting by $t \in \{0.6, 0.75, 0.9, 1.1, 1.25, 1.4\}$

as suggested in (Boddapati *et al.*, 2017) and we have found that more than 25% signal compression ($t < 0.75$) does not increase the F1 score of our approach. This makes sense because pitch-shifting with $t < 1$ is a lossy operation and it might increase the chance of losing pivotal frequency components. Overall, for pitch-shifting with $t < 1$ we kept only the two most influential values (0.75 and 0.9). For pitch-shifting with $t > 1$ we started with 1.1 and gradually increase it by 0.05 displacement to the margin of 65% signal stretching compared to the original signal. Stretching signals with $t > 1.65$ did not result in a positive effect on the performance of the front-end classifier. We speeded up all audio samples with $t \in \{1.1, 1.15, 1.2, 1.25, 1.3, 1.35, 1.4, 1.45, 1.5, 1.55, 1.6, 1.65\}$ and ranked them with respect to the F1 score measured for the front-end classifier. We finally kept $t \in \{1.15, 1.5\}$ for stretching the audio signal as the rest did not show any considerable improvement. Therefore, using static pitch shifting scales of 0.75, 0.9, 1.15, and 1.5, we ended up with an augmented dataset of 43 660, 10 000, 2 000 and 23 400 samples for UrbanSound8k, ESC-50, ESC-10 and DCASE-2017 datasets, respectively.

For each audio sample in the augmented datasets, we generate the DWT spectrograms by setting the sampling frequency of 8 kHz for ESC-10 and UrbanSound8k datasets, and 16 kHz to ESC-50 and DCASE-2017 datasets. Besides, we also set the frame length to 50 ms for ESC-10 and UrbanSound8k, 30 ms for ESC-50, and 40 ms for DCASE-2017 with a fixed overlapping size of 50% (Boddapati *et al.*, 2017). Therefore, each audio samples is now represented by a DWT spectrogram of 768×384 pixels. Empirically, this resolution provides a fair trade-off between information content (in terms of feature vectors) and dimensionality. Each spectrogram undergoes the enhancement step and next we apply the high-level data augmentation using the proposed WCCGAN.

The proposed WCCGAN employs the ConvNets presented in Figure 2.2, which have normalized convolution layers by applying the instance normalization (Ulyanov, Vedaldi & Lempitsky, 2016) technique followed by the leaky ReLU activation function with slope 0.3. We used the Glorot weight normalization algorithm for improving learning. For discriminator functions D_T and D_S we use a single architecture as depicted in Figure 2.3. In the proposed architecture

Table 2.1 The total number of epochs

Dataset	# of Training Epochs			
	$F_{S\rightarrow T}$	$F_{T\rightarrow S}$	D_S	D_T
ESC-10	123	107	104	91
ESC-50	136	118	109	116
UrbanSound8k	112	106	97	45
DCASE-2017	213	143	90	102

we set the receptive field to 3×3 and strides are set to 1×1 and 2×2 for the first and second convolution layers, respectively. In this case, we used the ReLU activation function and batch normalization (Gulrajani, Ahmed, Arjovsky, Dumoulin & Courville, 2017). These four networks are trained in four parallel GPUs GTX580 based on an implementation proposed in (Krizhevsky *et al.*, 2012). We applied an early stopping policy for training these networks and the total number of epochs for training each network is shown in Table 2.1.

The tentative values for c_1, c_2, μ, σ, and α in Equations 2.8, 2.9, and 2.11 for each dataset are shown in Table 2.2. There is no deterministic approach to adjust such hyperparameters of the WCCGAN. Moreover, there is no guarantee that such hyperparameters are properly set, as they result from exploratory experiments where we empirically modified them up to see a good track in sample generation and detection. In all experiments the main criterion was achieving the best epoch before overtraining generators and discriminators using early stopping. We have changed the hyperparameters almost randomly to get the best epoch. Since $F_{S\rightarrow T}$ is stronger than $F_{T\rightarrow S}$ due to the residual blocks, we intentionally increase the weight of the latter generator for all the datasets. This is the main reason for having higher values for σ compared to μ in all experiments. Hyperparameters c_1 and c_2 are weights for source and target samples respectively. Hence, except for the DCASE-2107 dataset, we tried to keep the summation of these weights close to one to ensure good balance. The hyperparameter α keeps the cycle consistency and we noticed that it should not exceed 0.45 for the proposed setup as higher values do not lead to convergence of the generators. Table 2.2 shows the hyperparameter values found by a basic and non-optimal local random search that attempts to find the models that produce the best F1 score

Table 2.2 Hyperparameters for Eq. 2.8, 2.9, and 2.11

Dataset	Hyperparameter Values (averaged over five folds)				
	c_1	c_2	μ	σ	α
ESC-10	0.49	0.67	0.02	0.76	0.23
ESC-50	0.39	0.68	0.12	0.58	0.19
UrbanSound8k	0.62	0.36	0.14	0.57	0.03
DCASE-2017	0.03	0.21	0.18	0.43	0.31

in terms of a minimum number of epochs. Once the best hyperparameters have been found, we applied perturbations of ±2%, ±5%, and ±10% to assess the sensitivity of the WCCGAN in respect to these hyperparameters. The F1 score of the discriminator networks has been computed for each perturbation applied on the hyperparameters of Table 2.2 which resulted in a noticeable performance drop, ranging from 2.4% to 12.6%, depending on the dataset and hyperparameter. As expected, these hyperparameters have a great influence in the performance of the WCCGAN because they are tuned upon a local search to allow the WCCGAN to produce spectrograms with low inter-class and high intra-class similarity. Among all these hyperparameters, α is the most sensitive one as it controls the consistency. In other words, this hyperparameter leverages the cycle-consistency loss between generators and acts to some extent as a regularizer for the generators.

In order to produce more structural spectrograms from source S to target T and make the loss functions converge, we need to have an idea of the inter-class relation between samples. For such an aim, we randomly pick samples to train a RF algorithm on spectrograms without high-level data augmentation featuring different number of trees from 500 to 3 000. Table 2.3 shows the confusion matrices for the RF trained with the UrbanSound8k dataset without high-level data augmentation. The values in Table 2.3 can also be interpreted as similarity among classes. For instance, class "EN" has high similarity with class "AI" because the classifier has misclassified samples from the class "AI" as class "EN" in 14% of the cases. Therefore, we set the source and target classes in Figure 2.1 to S="AI" and T="EN", respectively. We use the same procedure for all classes. In addition to intra-class image-to-image translation, we augment the DWT

Table 2.3 Confusion matrix of the proposed classification approach without high-level augmentation on the UrbanSound8k dataset. Values in bold indicate the best recognition accuracy in a 5-fold cross validation setup.

	AI	CA	CH	DO	DR	EN	GU	JA	SI	SM
AI	**0.68**	0.00	0.02	0.01	0.05	0.14	0.01	0.04	0.02	0.03
CA	0.00	**0.77**	0.02	0.02	0.00	0.00	0.00	0.05	0.07	0.07
CH	0.07	0.05	**0.31**	0.09	0.04	0.03	0.02	0.04	0.15	0.20
DO	0.06	0.04	0.03	**0.68**	0.04	0.02	0.03	0.00	0.05	0.05
DR	0.02	0.04	0.02	0.02	**0.74**	0.01	0.01	0.10	0.04	0.00
EN	0.04	0.00	0.03	0.02	0.01	**0.78**	0.02	0.06	0.01	0.03
GU	0.00	0.02	0.00	0.03	0.00	0.00	**0.95**	0.00	0.00	0.00
JA	0.01	0.01	0.00	0.00	0.05	0.03	0.00	**0.90**	0.00	0.00
SI	0.03	0.06	0.03	0.02	0.02	0.01	0.03	0.01	**0.78**	0.01
SM	0.03	0.08	0.06	0.09	0.08	0.08	0.01	0.04	0.06	**0.47**

Table 2.4 Confusion matrix of the proposed classification approach with WCCGAN augmentation on the UrbanSound8k dataset. Values in bold indicate the best recognition accuracy in a 5-fold cross validation setup.

	AI	CA	CH	DO	DR	EN	GU	JA	SI	SM
AI	**0.89**	0.01	0.00	0.02	0.01	0.04	0.00	0.01	0.02	0.00
CA	0.01	**0.92**	0.00	0.01	0.02	0.00	0.01	0.01	0.00	0.02
CH	0.00	0.01	**0.91**	0.03	0.00	0.01	0.00	0.00	0.01	0.03
DO	0.00	0.00	0.01	**0.96**	0.01	0.00	0.00	0.00	0.01	0.01
DR	0.00	0.01	0.00	0.02	**0.95**	0.01	0.00	0.01	0.00	0.00
EN	0.01	0.00	0.00	0.01	0.00	**0.96**	0.01	0.00	0.00	0.01
GU	0.01	0.00	0.00	0.00	0.01	0.00	**0.97**	0.01	0.00	0.00
JA	0.02	0.00	0.00	0.00	0.00	0.00	0.00	**0.98**	0.00	0.00
SI	0.00	0.01	0.03	0.00	0.01	0.00	0.00	0.00	**0.95**	0.00
SM	0.00	0.01	0.02	0.01	0.00	0.00	0.01	0.00	0.00	**0.95**

spectrograms in inter-class manner as well. We randomly select 50% of samples within a class as the source and the remaining 50% as the target classes. Overall, we increase the size of the datasets with extra 1 500, 2 000, 5 000 and 4 500 samples for ESC-10, ESC-50, UrbanSound8k and DCASE-2017, respectively. Some visual examples of the generated spectrograms using the WCCGAN are shown in Figure 2.4. This figure shows the high capability of the WCCGAN for producing structurally similar spectrograms even when the source and target are not similar to each to the human eye perspective.

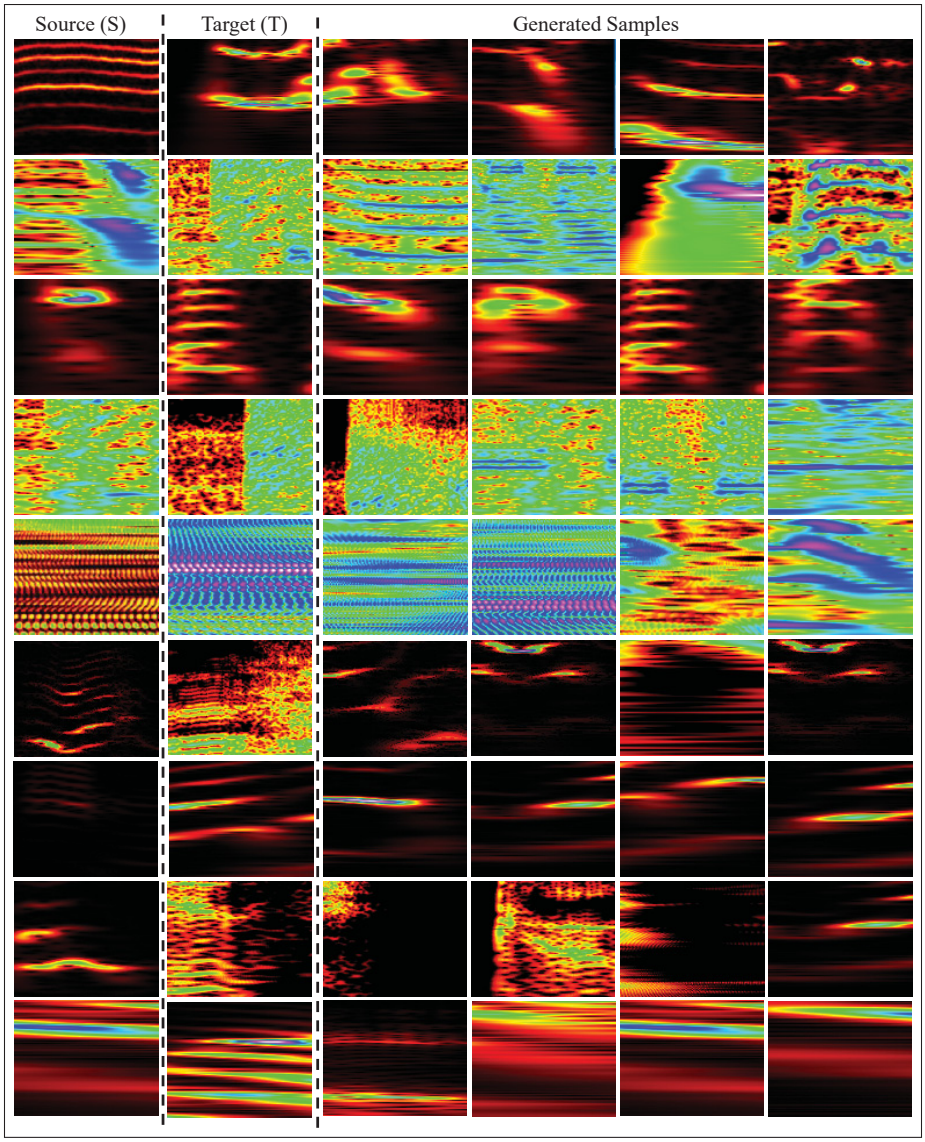

Figure 2.4 Generated spectrograms using the WCCGAN for randomly drawn sources (S) and targets (T). The Ss and Ts shown in the top four rows indicate intra-class image-to-image translation. Specifically, UrbanSound8k ($S = T$: sea waves), ESC-10 ($S = T$: person sneeze), ESC-50 ($S = T$: pouring water), and DCASE-2017 ($S = T$: office). Sources and targets for inter-class translation are shown in the five bottom rows as in UrbanSound8k (S: sea waves, T: rain), ESC-10 (S: person sneeze, T: helicopter), ESC-50 (S: wind, T: pouring water), and DCASE-2017 (S: cafe, T: office).

After finishing both inter- and intra-class data augmentation processes, we train again the RFs on the augmented dataset, considering different number of trees. The best number of trees for ESC-10, ESC-50, UrbanSound8k, and DCASE-2017 were obtained at 2 000, 1 864,

2 500, and 2 496, respectively with minimum AUC metrics (one-vs-all). Table 2.4 shows the performance of the learned trees on the UrbanSound8k dataset augmented with the proposed WCCGAN. The results are highly improved compared to the trees trained on codebooks learned without high-level data augmentation (Tables 2.3). This shows the importance of high-level data augmentation for extracting more discriminating features.

Table 2.5 compares the performance of the proposed classification approach to the state-of-the-art pre-trained classifiers (AlexNet and GoogLeNet) on environmental sound datasets following the transfer learning and fine-tuning strategies explained in (Kumar, Khadkevich & Fügen, 2018). It is worth mentioning that these two pre-trained networks have been fine-tuned on the 2D aggregation (pooling) of STFT, MFCC, and CRP. As Table 2.5 shows, our approach outperforms both deep learning models on all environmental sound datasets. One clear outcome of Table 2.5 is that the GAN theory could help us not only to build robust classifiers, but also to highlight another traditional classifier's performance. Furthermore, for a better comparison of the performances, the box-plots of these classifiers are shown in Figure 2.5. With respect to these box-plots for all the four benchmarking datasets, the proposed approach using the WCCGAN architecture for high-level data augmentation achieved the highest maximum, mean, minimum, and median accuracy. These plots also confirm that the proposed approach together with WCCGAN outperforms AlexNet and GoogLeNet since it provides the highest statistical measures except for the ESC-50 dataset; and there are no outliers. In order to investigate the statistical significance of the recognition performances reported in Table 2.5, we used Friedman's test which is the non-parametric version of the one-way ANOVA with some limited repeated measures (Hogg & Ledolter, 1987). Upon 19 runs (degrees of freedom), we could reach the p-value of 0.05 on average, which shows the high performance of the proposed approach.

Even if the current state-of-the-art is based on pre-trained ConvNets fine-tuned on the 2D aggregation of STFT, MFCC, and CRP, for a fair comparison as well as to evaluate the potential of the proposed WCCGAN to generate DWT spectrograms that may also improve the performance of other classification approaches, we have evaluated the performance of GoogLeNet and AlexNet on the DWT spectrograms augmented by the proposed WCCGAN. We fine-tuned these two

Table 2.5 Comparing the mean accuracy of the proposed approach with and without high-level augmentation (DA) with GoogLeNet and AlexNet. Comparison has been made in a 5-fold cross validation setup. The best results are shown in bold faces.

Benchmarking Dataset	Mean Accuracy			
			Proposed Approach	
	GoogLeNet	AlexNet	Without DA	With DA
ESC-10	0.83	0.83	0.72	**0.87**
ESC-50	0.71	0.64	0.55	**0.77**
UrbanSound8k	0.91	0.90	0.73	**0.94**
DCASE-2017	0.64	0.62	0.66	**0.76**

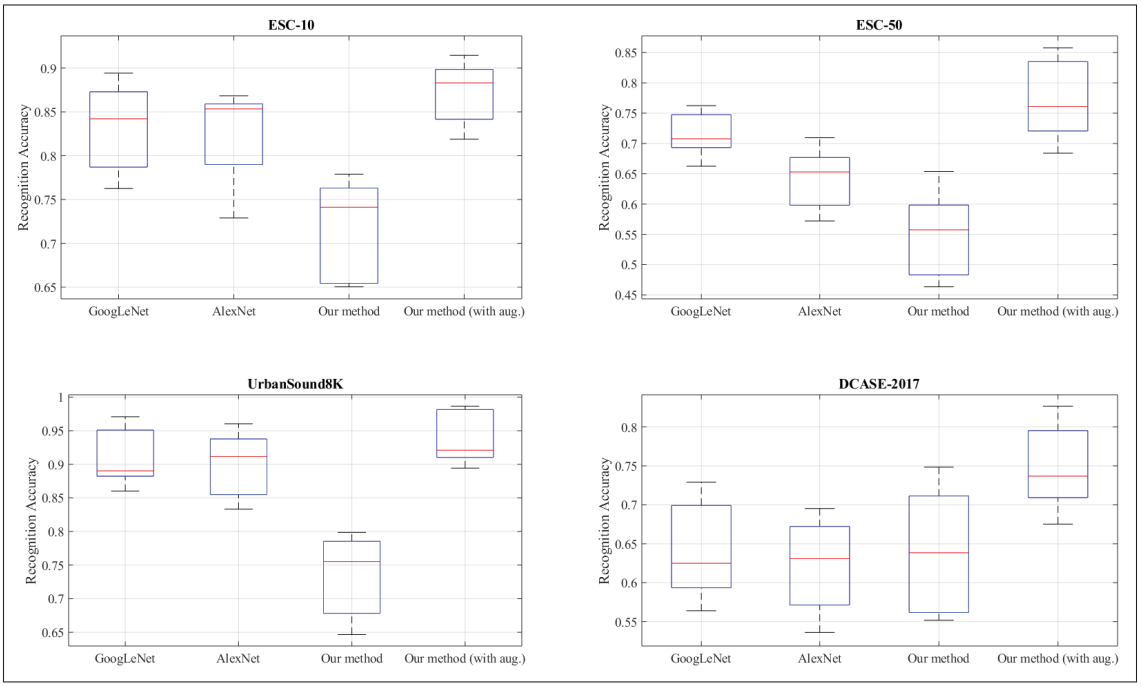

Figure 2.5 Box-plots of the approaches from Table 2.5 in a 5-fold cross validation setup for ESC-10, ESC-50, UrbanSound8k and DCASE-2017 datasets

ConvNets with the four augmented datasets and the results are shown in Table 2.6. The results show the importance of high-level data augmentation for environmental sound classification since the performance of these two ConvNets is also improved. With respect to the values reported in Table 2.6, GoogLeNet trained on the augmented DWT spectrogram outperforms the

Table 2.6 Recognition accuracy of two ConvNets on the augmented DWT spectrograms of four benchmarking datasets. Value in bold indicates a better performance than those reported in Table 2.5. The 5-fold cross validation setup is applied. The mean confidence refers to the probabilities computed by the softmax layer.

Benchmarking Dataset	Mean Accuracy (avg. AUC score)		Mean Confidence (%)
	GoogLeNet	AlexNet	
ESC-10	0.86	0.85	78.26
ESC-50	**0.78**	0.75	80.52
UrbanSound8k	0.93	0.93	91.02
DCASE-2017	0.73	0.74	81.37

Table 2.7 Average ranking ($\bar{r}$) considering the best mean accuracy for the four datasets (Brazdil & Soares, 2000)

Approach	$\bar{r}$ (averaged)	Overall Rank (according to $\bar{r}$)
Proposed Approach (with DA)	1.25	1
Proposed Approach (without DA)	5.00	6
GoogLeNet (with DA)	1.50	2
GoogLeNet	4.50	4
AlexNet (with DA)	2.50	3
AlexNet	4.75	5

proposed classification method on the ESC-50 dataset. Moreover, the performance of these two ConvNets is very close to our classification scheme.

Table 2.7 summarizes the comparison between all approaches with and without the proposed data augmentation through an average ranking (Brazdil & Soares, 2000) according to the measured mean accuracy. The proposed approach with data augmentation has the best rank among all approaches, followed by the GoogLeNet and AlexNet with data augmentation, GoogLeNet, AlexNet, and the proposed approach without data augmentation. The most impressive improvement due to the proposed data augmentation is observed for the proposed approach which moves from the last (6th) to the top spot (1st).

Table 2.8 Comparison of 1D data augmentations approaches in terms of recognition accuracy for the proposed classification scenario in a 5-fold cross validation setup. Note that after these 1D data augmentations, we have also augmented the DWT representations with WCCGAN.

Dataset	Mean Accuracy	
	All 1D Augmentation (low-level augmentation)	Pitch-shifting
ESC-10	0.79	0.87
ESC-50	0.75	0.77
UrbanSound8k	0.92	0.94
DCASE-2017	0.69	0.76

2.6 Discussion

We have shown the potential of the proposed WCCGAN for high-level data augmentation in improving the performance of two different supervised approaches (ConvNets and RFs). Since the proposed WCCGAN also considers inter-class and intra-class aspects to generate new samples, it allows generating more discriminating features as it improves recognition accuracy of all classifiers. Implementing low-level 1D data augmentation approaches proposed by Salamon et al. (Salamon & Bello, 2017) do not noticeably help to learn more informative features. Table 2.8 compares the results of several low-level 1D data augmentation approaches and a single low-level 1D data augmentation approach. For instance, we augmented the environmental datasets using all 1D augmentation approaches defined in (Salamon & Bello, 2017): time-stretching with scale of 0.81, 0.93, 1.07, and 1.23; pitch-shifting with factors of 0.75, 0.9, 1.15, and 1.5 (the same parameters defined in Section 2.5); dynamic range compression using three parameterizations from the Dolby E standard and one from the Icecast radio streaming server; and background noise using acoustic scenes of street-workers, street traffic, street-people, and park. Table 2.8 shows that employing all types of low-level data augmentation do not necessarily improve the performance of the classifier.

We have also compared the performance of the proposed WCCGAN with the CCGAN proposed by Zhu et al. (Zhu *et al.*, 2018c) on the DWT spectrograms. The input size of the generator and discriminator networks in the CCGAN is 48×48 which is considerably smaller than our

Table 2.9 Recognition accuracy of the proposed approach with different cycle-GAN augmentation architectures on DWT spectrograms. The 5-fold cross validation setup is applied and the bold values indicate the best performance.

Benchmarking Dataset	Mean Accuracy		
	CCGAN (Zhu *et al.*, 2018c)		WCCGAN (the proposed)
	(48×48)	(768×384)	(768×384)
ESC-10	0.74	0.75	**0.87**
ESC-50	0.67	0.70	**0.77**
UrbanSound8k	0.80	0.80	**0.94**
DCASE-2017	0.67	0.71	**0.76**

Table 2.10 Recognition accuracy of the ConvNet (Zhu *et al.*, 2018c) with different cycle-GAN augmentation architectures on DWT spectrograms. The 5-fold cross validation setup is applied and the bold values indicate the best performance.

benchmarking Dataset	Mean Accuracy			
	CCGAN (Zhu *et al.*, 2018c)		WCCGAN (The proposed)	
	(48×48)	(768×384)	(48×48)	(768×384)
ESC-10	0.40	0.41	0.59	**0.67**
ESC-50	0.41	0.44	0.51	**0.59**
UrbanSound8k	0.38	0.41	0.59	**0.64**
DCASE-2017	0.39	0.42	0.50	**0.52**

spectrogram dimensions of 768×384. For a fair comparison, we have adapted the input dimensions of the networks to the size of our generated spectrograms as well as we have squeezed the DWT spectrograms to 48×48 to fit them to the networks. The results of these experiments are summarized in Tables 2.9 and 2.10. Table 2.9 shows that the proposed WCCGAN outperforms the architecture introduced in (Zhu *et al.*, 2018c) considering our front-end RF classifier for both input dimensions. Table 2.10 shows that the proposed approach also outperforms the CCGAN when we use the ConvNet proposed by Zhu et al. (Zhu *et al.*, 2018c) as a front-end classifier. These results show the advantage of the proposed WCCGAN and front-end classification compared to the classification pipeline proposed in (Zhu *et al.*, 2018c) for spectrograms.

Table 2.11 Mean accuracy of different environmental sound classification approaches in UrbanSound8k (US8K), ESC-10, ESC-50 and DCASE-2017 datasets with and without data augmentation (DA). Values are rounded in two-digit floating point precision.

Approach	Mean Accuracy			
	US8K	ESC-10	ESC-50	DCASE-2017
Proposed Approach (DA)	0.94	0.87	**0.77**	**0.76**
MC-Net + LMC-Net (Su *et al.*, 2019)	**0.95**	0.72	0.74	0.74
GooLeNet and AlexNet (Boddapati *et al.*, 2017)	0.93	0.86	0.73	NA
SoundNet (Aytar *et al.*, 2016)	0.79	**0.92**	0.74	NA
SB-ConvNets (DA) (Salamon & Bello, 2017)	0.79	0.77	0.54	0.45
MoE (Ye, Kobayashi & Murakawa, 2017)	0.77	NA	NA	NA
SKM (DA) (Salamon & Bello, 2015b)	0.76	0.74	0.56	0.43
SKM (Salamon & Bello, 2015b)	0.74	0.71	0.52	0.36
Proposed Approach	0.73	0.71	0.55	0.66
PiczakConvNets (Piczak, 2015a)	0.73	0.80	0.65	0.52
SB-ConvNets (Salamon & Bello, 2017)	0.73	0.72	0.49	0.41
MultiTemp (Zhu, Xu, Wang, Zhang, Li & Peng, 2018a)	0.72	0.74	0.71	0.73
VGG (Pons & Serra, 2019)	0.70	NA	NA	NA

NA: Not Available.

Finally, Table 2.11 shows the mean classification accuracy of the proposed approach with and without data augmentation as well as the results obtained by other state-of-the-art classifiers described in the literature. The proposed approach achieved the highest mean accuracy for ESC-50 and DCASE-2017 and its performance is just 0.01 lower than the approach based on the decision-level fusion of two parallel ConvNets (MC-Net + LMC-Net) for the UrbanSound8k dataset. However, the best performance for the ESC-10 dataset is achieved by the Soundnet (Aytar *et al.*, 2016) which learns a multimodal representation from a very-large dataset of unlabeled videos which is further used with an SVM. Besides, the proposed approach outperforms most of the approaches trained on handcrafted features or trained on both 1D signal and spectrograms.

2.7 Conclusion

In this paper we have shown how to structurally augment imbalanced environmental sound datasets in a high-level fashion using the proposed WCCGAN. The proposed data augmentation

framework applies identity mapping to discriminator networks, which using the least-squared optimization criterion solves the gradient vanishing problem and produces flawless spectrograms. The importance of the high-level augmentation is more tangible for spectrograms because compared to regular computer vision datasets (e.g., ImageNet (Deng, Dong, Socher, Li, Li & Fei-Fei, 2009)), spectrograms do not have solid objects sensitive to low-level transformations. Moreover, the total number of samples in environmental sound datasets are limited and image-to-image translation using the WCCGAN can effectively increase the size and improve the quality of the datasets. The proposed high-level data augmentation approach is also able to produce consistent samples that keep structural significance which is much more meaningful compared to other approaches such as simple image transformations or even conventional GANs. Such approaches do not allow control of the generated samples, especially regarding their structural consistency. The experimental results have shown that the WCCGAN outperforms the regular GAN since we do not have much control over consistency of the source and target inputs. Overall, high-level data augmentation using GANs translates structural components from sample to sample where low-level augmentation algorithms cannot. Furthermore, the experimental results have also shown that the WCCGAN can even improve the performance of ConvNets for the environmental sound classification task. Unfortunately, the proposed architecture for the cycle-consistent GAN does not properly work in an end-to-end 1D setup. In fact, it is really costly to train and find hyperparameters for an end-to-end WCCGAN as audio waveforms have a much higher dimensionality compared to spectrograms. In spite of the high dependence of the proposed architecture on the dataset, we believe that the proposed WCCGAN can also be adapted to other datasets with some customization in the architecture of generators and discriminators and an appropriate hyperparameter tuning. The burden of hyperparameter tuning may be reduced by using a black-box optimization such as the Ortho-MAD2S (Mello, de Matos, Stemmer, Britto Jr. & Koerich, 2019).

Our classification approach is a promising step towards building reliable classifiers for complex environmental sound datasets. We learn a codebook with visual words extracted by SURF detectors from augmented spectrograms organized in a unit distance to each other in a setup

imposed by the K-Means++ algorithm. Unsupervised feature learning has shown great competence in classifying 2D representations of the environmental sound datasets. The RF classifier with 2 000 trees trained on code vectors outperformed the two ConvNets in four benchmarking datasets (ESC-10, ESC-50, UrbanSound8k, and DCASE-2017). Furthermore, besides outperforming deep models, the unsupervised feature learning approach together with the proposed architecture for the WCCGAN compares favorably with most of the current approaches for environmental sound classification. Another aspect is the reliability of the proposed approach against adversarial attacks. It is out of the scope of this paper to discuss this aspect, but it has been shown that ConvNets such as AlexNet and GoogLeNet are more vulnerable against carefully crafted adversarial examples compared to classifiers trained with SURF feature vectors (Esmaeilpour *et al.*, 2020).

For the future work, in addition to improving the Spherical K-Means++ algorithm for environmental sound classification, we would like to measure the performance of other unsupervised algorithms on the augmented DWT datasets to understand the strength of these classifiers. Besides that we are also interested in evaluating Wasserstein GAN (Arjovsky *et al.*, 2017) for image-to-image translation since it suffers less from oversmoothing effects. This might improve further the performance of the proposed classification approach. Finally, we would like to extend this work for structured datasets such as music datasets and evaluate the performance of the proposed data augmentation and classification approaches.

CHAPTER 3

A ROBUST APPROACH FOR SECURING AUDIO CLASSIFICATION AGAINST ADVERSARIAL ATTACKS

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in « IEEE Transactions on Information Forensics and Security » in 2019.

Abstract

Adversarial audio attacks can be considered as a small perturbation unperceptive to human ears that is intentionally added to an audio signal and causes a machine learning model to make mistakes. This poses a security concern about the safety of machine learning models since the adversarial attacks can fool such models toward the wrong predictions. In this paper we first review some strong adversarial attacks that may affect both audio signals and their 2D representations and evaluate the resiliency of deep learning models and support vector machines (SVM) trained on 2D audio representations such as short time Fourier transform, discrete wavelet transform (DWT) and cross recurrent plot against several state-of-the-art adversarial attacks. Next, we propose a novel approach based on pre-processed DWT representation of audio signals and SVM to secure audio systems against adversarial attacks. The proposed architecture has several preprocessing modules for generating and enhancing spectrograms including dimension reduction and smoothing. We extract features from small patches of the spectrograms using the speeded up robust feature (SURF) algorithm which are further used to transform into cluster distance distribution using the K-Means++ algorithm. Finally, SURF-generated vectors are encoded by this codebook and the resulting codewords are used for training a SVM. All these steps yield to a novel approach for audio classification that provides a good trade-off between accuracy and resilience. Experimental results on three environmental sound datasets show the

competitive performance of the proposed approach compared to the deep neural networks both in terms of accuracy and robustness against strong adversarial attacks.

3.1 Introduction

Adversarial attacks pose security issues since they can be unrecognizable to human eyes or human ears while they can easily fool any trained machine learning model with very high confidence. As these machine learning models are becoming more present in many devices and applications, there exists an urgent need for improving their robustness against adversarial attacks. Basically, an adversarial attack algorithm formulates an optimization problem such as finding the smallest possible perturbation to be added to a given legitimate input (image, audio, spectrogram, etc.) aiming at a machine learning model to predict a wrong label. This perturbation should be as small as possible to be imperceptible to human visual or auditory system. Adversarial attacks have been attracting the attention of many researchers, mainly in the domain of computer vision (Kurakin, Goodfellow & Bengio, 2016; Sabour, Cao, Faghri & Fleet, 2015; Xie, Wang, Zhang, Zhou, Xie & Yuille, 2017). However, adversarial attacks may also pose a serious threat to voice assistant devices, speech and speaker recognition as well as other audio-related applications. In spite of that few studies have addressed adversarial attacks for audio signals (Carlini & Wagner, 2018). One of the possible reasons is the considerable optimization overhead of adversarial algorithms when applied to audio signals due to their high dimensionality. In the big picture, adversarial examples of audio signals can be crafted during sound production or post production by changing their amplitude or frequency into the ranges where humans cannot perceive. This is difficult and needs to be treated carefully because there is no guarantee of producing a true adversarial example and the output could be just a noisy example. In the case of post-production of adversarial examples, the adversary can either solve an optimization problem (costly) or develop an adversarial filter in order to apply some perturbations to a legitimate audio before passing it through a machine learning model. In both cases, the victim model could be fooled toward the bad wishes of the adversary and make the system misbehave.

In this paper, we investigate the threat of adversarial attacks on environmental audio sounds due to the diversity that we may find, ranging from baby crying to engines, horns to dog barking or people chatting with numerical text-free labels. Adversarial attacks are quite useful for other relevant domains of speech recognition and music classification and they may be generalizable to speech-to-text applications, though the latter is not discussed in this paper. Environmental sound classification has been a challenging problem in machine learning research (Salamon & Bello, 2015b). Both shallow and deep neural networks (DNNs) have shown competitive performances on benchmarking datasets such as ESC-10 (Piczak, 2015b), ESC-50 (Piczak, 2015b), and UrbanSound8K (Salamon *et al.*, 2014a). Besides the supervised models, there are some unsupervised models such as spherical K-means for sound representation learning (Salamon & Bello, 2015b,a). Both supervised and unsupervised models have mainly been trained either on audio waveforms (1D) or on 2D representation such as spectrograms. In both cases, convolutional neural networks (CNNs) have shown better performances compared to other classifiers. For instance, the CNN proposed by Salamon and Bello (Salamon & Bello, 2017) outperforms their prior approach based on unsupervised feature learning and random forest (Salamon & Bello, 2015b) on the UrbanSound8K dataset. Also, for ESC-10 and ESC-50 datasets, a 1D CNN with eight convolution layers (SoundNet) (Aytar *et al.*, 2016) outperforms random forest (Piczak, 2015b), SVM using Mel-Frequency Cepstral Coefficients (MFCCs) (Piczak, 2015b), and convolutional autoencoders (Aytar *et al.*, 2016). In addition to these CNNs, other DNN architectures such as AlexNet and GoogLeNet, which have shown remarkable performances on image classification tasks (e.g. ImageNet dataset) have also been used for environmental sound classification. Interestingly, these two CNNs trained on spectrograms have been achieving the highest recognition performance for the three aforementioned datasets as reported by Boddapati *et al.* (Boddapati *et al.*, 2017).

One of the open problems in audio classification seemingly is no longer improving recognition accuracy but improving their strengths against some carefully crafted adversarial examples. Therefore, the proposed approach for environmental sound classification is based on two findings: (i) deep learning models, particularly AlexNet and GoogLeNet, outperform conventional

classifiers trained on handcrafted features such as SVM; (ii) SVM in general is more robust against adversarial attacks, potentially because it learns from low-dimensional feature vectors that might reduce the chance of being affected by adversarial perturbations compared to deep models which learn from raw data. Following these facts, in this paper we propose an SVM-based approach that provides a good trade-off between the recognition accuracy and the robustness against adversarial attacks while achieving recognition accuracy comparable to deep models. Since there is no standard metric for evaluating the quality of such a trade-off, we also introduce a distance metric based on the error rate versus the fooling rate.

Our contribution in this paper is threefold: (i) we present common adversarial attacks for audio and we show how they can affect the security of audio applications; (ii) we characterize the vulnerability of state-of-the-art models based on 2D representations to adversarial attacks and the transferability of these attacks between different models; (iii) we propose a novel approach for environmental sound classification that in addition to being robust against several adversarial attacks without incorporating any reactive or proactive defense process, also provides a high recognition accuracy, which is competitive with the state-of-the-art.

This paper is organized as follows. Section 3.2 introduces general adversarial attacks and describes the most important ones. In this section we also present the adversarial attacks that may affect audio applications based on 2D audio representations and discuss adversarial attacks that may affect audio waveforms. Section 3.3 presents the main 2D representations for audio signals. Section 3.4 presents the proposed approach that aims to achieve both good classification accuracy and robustness to adversarial attacks. In Section 3.5 we characterize the vulnerability of some state-of-the-art models in the problem of environmental sound classification, measure the resiliency of the proposed approach versus CNNs and review the adversarial example transferability among these models. The conclusions and perspectives of future work are presented in the last section.

3.2 Adversarial Attacks

Adversarial attacks can be considered as carefully crafted perturbations that when intentionally added to a legitimate example, lead machine learning models to misbehave (Weng, Zhang, Chen, Yi, Su, Gao, Hsieh & Daniel, 2018). Considering $\mathbf{x}$ as a legitimate example, then an adversarial example $\mathbf{x}'$ can be crafted in such a way that:

$$\mathbf{x} \approx \mathbf{x}', \qquad f^*(\mathbf{x}) \neq f^*(\mathbf{x}') \tag{3.1}$$

where f^* is the post-activation function. Supposing that $\mathbf{x}$ represents an image or an audio signal, the differences between $\mathbf{x}$ and $\mathbf{x}'$ should not be perceived by the human visual or auditory systems.

There are several algorithms for generating $\mathbf{x}'$, mainly when $\mathbf{x}$ is an image. The adversarial attacks can be categorized into different groups. For instance, if the adversary has access to the model architecture, parameters, training dataset, etc., it is categorized as a white-box attack, otherwise it is called black-box. Also, adversarial attacks can have other taxonomy such as targeted, where the adversarial perturbation is crafted having in mind a specific target label, and non-targeted, where the adversarial perturbation is crafted to induce a machine learning model to predict any incorrect label. Due to the importance of studying adversarial threats for data-driven machine learning models, many attack algorithms have been proposed and they have shown a great success in fooling advanced models. However, the main challenge of almost all attack algorithms is their computational complexity, which makes adversarial training very time-consuming.

One of the first proposed attacks is the Fast Gradient Sign Method (FGSM) (Goodfellow *et al.*, 2015), which still remains one of the most effective attacks. FGSM was originally built to attack CNNs but it can also be a serious threat for non-deep architectures. FGSM generates an adversarial example $\mathbf{x}'$ by:

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{w}, \mathbf{x}, y)) \tag{3.2}$$

where $\mathbf{x}$ and y are the legitimate input and its true label respectively, ϵ is a constant value which can be determined by an optimization scheme, and J is the cost function for the model parameter $\mathbf{w}$ obtained after completing the training process. FGSM is a white-box attack which means that the model parameter $\mathbf{w}$ should be accessible to fetch its gradient information and generate the adversarial example $\mathbf{x}'$. In other words, by providing the trained model and the training dataset, FGSM can generate adversarial examples $\mathbf{x}'$ using Eq. 3.2, which have unrecognizable differences to the legitimate input $\mathbf{x}$ and $\mathbf{x}'$ can perhaps make the model $\mathbf{w}$ to predict a wrong label $y' \neq y$ with high confidence.

The iterative version of the FGSM attack is known as Basic Iterative Method (BIM) (Kurakin *et al.*, 2016) and its attack frequency (i.e., the number of iterations in running an attack) is higher than one. In fact, BIM's optimization procedure can stop after generating the first adversarial example (BIM-a) or continue up to a pre-defined number of iterations (BIM-b). These two attacks are actually the improved version of FGSM which increases the attack rate, at the cost of higher computational complexity.

Carlini and Wagner (Carlini & Wagner, 2017b) have proposed an optimization-based attack known as CWA, which uses the similarity metric d_i defined in Eq. 3.3.

$$d_i = \left\| \mathbf{x}_i - \mathbf{x}'_i \right\| \tag{3.3}$$

where i is the sample index. CWA attempts to minimize d_i as:

$$\min_c \|d_i\| + c \times g(\mathbf{x} + d_i) \quad \text{s.t.} \quad \mathbf{x} + d_i \in [0, 1]^n \tag{3.4}$$

where $c > 0$ is a suitably chosen constant, $g(d) \geq 0 \iff f(d) = y'$; and y' is the wrong label for $\mathbf{x}$. The intuition behind Eq. 3.4 is similar to the dropout variational inference introduced by Li *et al.* (Li & Gal, 2017). This attack is very similar to the FGSM attack with two main differences: (i) it changes the input $\mathbf{x}_i$ using the tanh function; (ii) it uses a difference between logits (the vector of non-normalized predictions that a model generates) instead of optimizing a cost function for regular cross-entropy. CWA is one of the strongest iterative and targeted

adversarial attacks and it can be very effective in fooling CNNs, though costly as it might need too many callbacks Carlini & Wagner (2017b) to $\mathbf{x}$.

The adversarial attacks presented so far are designed for DNNs. Since the approach proposed in this paper is based on SVMs, we also present two adversarial attacks designed to attack SVM models: Evasion attack (EA) and Label Flipping attack (LFA). EA (Biggio, Corona, Maiorca, Nelson, Šrndić, Laskov, Giacinto & Roli, 2013) and LFA (Xiao *et al.*, 2012). The main difference between these two attacks is that LFA contaminates the training data by flipping the true labels of the samples, while EA manipulates the sample distribution aiming to change the true labels. In both cases, the decision boundary of the model is shifted toward maximum loss for the test set. The general intuition behind EA is to map an input $\mathbf{x}$ over a support vector(s) by simply flipping its label. This flipping can be toward the trained weight direction(s) of the SVM as given by Eq. 3.5.

$$\mathbf{x}' = \mathbf{x} - \epsilon \odot \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|} \tag{3.5}$$

where $\mathbf{x}'$ is the crafted adversarial example, $\mathbf{w}_i$ is the weight vector discriminating support vectors, and ϵ is a small constant value. The intuition behind these two attacks is the geometrical definition of support vectors as given by Eq. 3.6.

$$\min \mathbf{w} \qquad \text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 \quad i = 1, \ldots, n \tag{3.6}$$

where $\mathbf{w}$ is a vector normal to the hyperplane ($\mathbf{w}^\top \mathbf{x} - b = 0$), b is a bias term, and $y = \{+1, -1\}$ is the label. The position of the support vectors can be depicted as shown in Fig. 3.1.

In other words, the SVM model will be fooled by moving a datapoint perpendicularly toward the opposite direction of its weight vector. This attack is generalizable to soft margin SVM by simply optimizing the value of ϵ in Eq. 3.7.

$$\frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i(\mathbf{w}^\top \mathbf{x} - b), 0) + \epsilon \|\mathbf{w}\|^2 \tag{3.7}$$

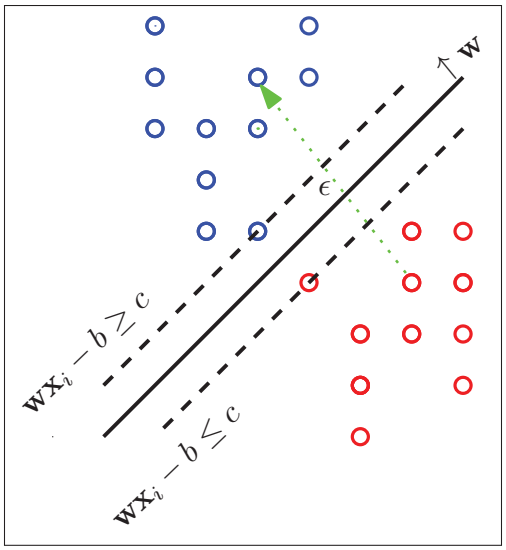

Figure 3.1 Visualization for Eq. 3.10

As long as the optimization of ϵ is perpendicularly directed toward the $\mathbf{w}_i$, the SVM model cannot distinguish an adversarial from legitimate examples. This data contamination in EA can be implemented by taking advantage of gradient information and local search for achieving the best data perturbation with a specific budget as introduced by Biggio *et al.* (Biggio *et al.*, 2013). The gradient information can be exploited for different kernels. For an RBF kernel with variance σ^2, we have $K(\mathbf{x}, \mathbf{x}_i) = \exp(-0.5 \cdot \sigma^{-2} \|\mathbf{x} - \mathbf{x}_i\|^2)$, and its gradient can be computed by Eq. 3.8.

$$\nabla K(\mathbf{x}, \mathbf{x}_i) = -\sigma^{-2} \exp(-0.5 \cdot \sigma^{-2} \|\mathbf{x} - \mathbf{x}_i\|^2)(\mathbf{x} - \mathbf{x}_i) \tag{3.8}$$

Similarly, for a polynomial kernel of degree p, denoted as $K(\mathbf{x}, \mathbf{x}_i)=(\langle \mathbf{x}, \mathbf{x}_i \rangle + c)^p$, its gradient can be computed by Eq. 3.9.

$$\nabla K(\mathbf{x}, \mathbf{x}_i) = p(\langle \mathbf{x}, \mathbf{x}_i \rangle + c)^{p-1} \mathbf{x}_i \tag{3.9}$$

Therefore, the adversarial example $\mathbf{x}'$ can be computed by:

$$\mathbf{x}' = \mathbf{x} - \eta \nabla f(\mathbf{x}) \tag{3.10}$$

where η is a small scalar (step size) and f denotes the learned filters in the hypothesis space H for $K(\mathbf{x}, \mathbf{x}_i)$=$\Phi(\mathbf{x})^\top \Phi(\mathbf{x}_i)$ and Φ is a mapping function from input to the feature space. Unlike EA, LFA does not generate an adversarial example via distorting the legitimate samples, but it contaminates the labels of such samples. This should result in maximum loss in the test set while it is expected to be minimum for the training set. The LFA attack can be implemented by solving the following optimization problem:

$$\min_{q,\mathbf{w},\epsilon,b} \gamma \sum_{i=1}^{2n} q_i(\epsilon_i - \xi_i) + \frac{1}{2} \|\mathbf{w}\|^2 \tag{3.11}$$

subject to:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \epsilon_i \quad \epsilon_i \geq 0, \quad i = 1, \cdots, 2n \tag{3.12}$$

having the budget of:

$$\sum_{i=n+1}^{2n} c_i q_i \leq C \tag{3.13}$$

where γ is a fixed positive parameter for quantifying the trade off, $q_i \in \{0, 1\}$ is an indicator variable for controlling over the legitimate (q=0) and contaminated example (q=1), and c_i and C denote the flipping cost of each example and the total flipping cost, respectively, from adversary's point of view. The hinge loss function ($\mathcal{L}$), defined in Eq. 3.14

$$\mathcal{L}(y, f(\mathbf{x})) := \max(0, 1 - y f_D(\mathbf{x})) \tag{3.14}$$

This loss function has been also used for ϵ_i on the contaminated dataset of D' as:

$$\epsilon_i := \max(0, 1 - y_i f_{D'}(\mathbf{x}_i)) \tag{3.15}$$

where D' is the contaminated dataset which also includes the original dataset D. Similarly, ξ_i refers to the hinge loss of the classifier f_D:

$$f_D(\mathbf{x}) := \mathbf{w}^\top \mathbf{x} + b, \quad \mathbf{w} := \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i). \tag{3.16}$$

Herein, b is the bias term and α denotes the Mercer kernel coefficient of the SVM.

3.2.1 Transferability of Adversarial Attacks

One of the main characteristics of adversarial attacks described so far is that they are non-targeted toward a specific label as they maximize the probability of any label other than the true one. This is very tricky since it opens up the opportunity of adversarial transferability to other data-driven models. This means that adversarial examples maintain their effectiveness against models different from those targeted by the attack. For instance, the FGSM attack, which targets CNNs, could completely fool a maxout network trained on the MNIST dataset (Goodfellow *et al.*, 2015). Goodfellow *et al.* (Goodfellow *et al.*, 2015) have shown that the linear behavior of FGSM can be transferred to other classifiers including SVMs even with radial basis kernel function. This was a breakpoint of studying adversarial transferability for all classifiers, from logistic regression (simple) to very-deep CNNs (complex). Recently, Sabour *et al.* (Sabour *et al.*, 2015) have shown the great effectiveness of FGSM on fooling other deep architectures with and without convolution layers.

A lot of effort has been made on improving transferability of adversarial attacks. From expanding input patterns (data-wise) (Xie, Zhang, Wang, Zhou, Ren & Yuille, 2018) to developing ensemble models that produce more misleading adversarial examples (model-wise) (Liu, Chen, Liu & Song, 2016). Therefore, this is a real threat since adversarial attacks can be transferred among almost all models, e.g. from CNN to SVM, logistic regression and decision trees (Papernot, McDaniel & Goodfellow, 2016c). Besides that models trained for speech-to-text translation have also been successfully fooled by crafted adversarial examples (Carlini & Wagner, 2018). Empirically, machine learning models designed for audio applications, based on either 1D or 2D representation are very vulnerable against adversarial attacks and the current defense schemes, such as those proposed by Das *et al.* (Das *et al.*, 2018), do not work appropriately.

3.2.2 Adversarial Attacks for Audio Signals

Adversarial attacks have been mainly studied in the domain of computer vision to perturb images. It has been shown that 2D CNNs are quite vulnerable against white-box and black-box optimization-based attacks (Goodfellow *et al.*, 2015). However, these optimization-based attacks are usually very costly, and they require too many callbacks to each legitimate example, pixel-by-pixel. Generalizing these optimization-based attacks to audio signals (1D) is not straightforward since the audio signal is usually high-dimensional data, even considering a single audio channel. For instance, five seconds of mid-quality audio corresponds to an array of 110,250 points. Therefore, computing a similarity measure such as the ℓ_2-norm between legitimate and crafted examples as a part of an adversarial optimization criterion is very challenging compared to 2D arrays.

Alzantot *et al.* (Alzantot, Balaji & Srivastava, 2018) and Du *et al.* (Du, Ji, Li, Gu, Wang & Beyah, 2019) have proposed speech-to-text adversarial attacks where the optimization process is replaced with heuristic algorithms like genetic algorithms (Alzantot *et al.*, 2018) or particle swarm optimization (Du *et al.*, 2019) to mitigate the considerable cost of the optimization process. Basically, these greedy and evolutionary algorithms introduce random noise to a legitimate example which in turn increases the chance of having a dissimilarity between legitimate and crafted adversarial examples. However, this also paves the way for an easy detection of adversarial examples by simple algorithms. On the other hand, in the most effective adversarial attacks for images (e.g. FGSM, BIM, CWA, etc.), adversarial perturbations are generated by an optimization process that has two key constraints: (i) induce a machine learning model to produce a wrong label; (ii) have a visual similarity between legitimate and adversarial examples.

It is difficult to satisfy these constraints for adversarial audio because it is very challenging and time-consuming optimizing for these two constraints considering the high dimensionality of audio signals. Moreover, in contrast with images, audio signals are not convolved in rows and columns and this also makes very difficult solving the optimization problem for adversarial audio perturbations. These difficulties constitute enough ground for introducing evolutionary

algorithms to randomly search for possible adversarial perturbations which basically can only respect the first key constraint. The main side effect of this approach is producing adversarial examples that stay close to the manifold of legitimate samples that can be easily detected by a tuned classifier or by a simple adversarial detector such as downsampling or upsampling. In this case, adversarial examples crafted by greedy algorithms lie in the submanifolds close to the legitimate samples, which is basically the same manifolds where noisy samples lie in.

Some adversarial attacks explicitly add noise to the audio signals mainly by manipulating the frequency components (Roy, Hassanieh & Roy Choudhury, 2017; Song & Mittal, 2017). Backdoor attack (Roy *et al.*, 2017) is based on adding non-linearity to an audio signal in frequency ranges inaudible to the human auditory system (over 20 kHz). This non-linearity can be captured by microphones but does not show recognizable effects on human ends. Taking advantage of this type of attack, perturbations can be computed in frequency domain and then applied to an audio signal, which can fool a machine learning model. Backdoor attack lacks in defining a general optimization formulation for computing adversarial frequency perturbations (the shadow signal) (Roy *et al.*, 2017). In other words, there is no analytical way for computing the perturbation. The potential perturbation value may change depending on the audio signal and therefore it makes the computation of proper shadow signals very cumbersome and time-consuming. Moreover, audio frequency manipulation, even if unrecognizable by humans, can be easily detected if the perturbed audio signal is converted to a 2D representation. For instance, adversarial examples generated by the Backdoor attack can be easily detected by a simple post-processing module which analyzes their spectrograms. An ideal case for an adversarial audio example is to be unrecognizable in both 1D and 2D representations. Similarly, DolphinAttack (Song & Mittal, 2017) implements phase domain manipulations to change the sample label toward other than the legitimate one that is unrecognizable by the human auditory system.

The detectability of the adversarial examples generated by Backdoor and DolphinAttack algorithms can be assessed by computing the local intrinsic dimensionality score (LID) (Ma *et al.*, 2018) for their 2D representations. For such an aim, three groups of inputs should be defined: normal, noisy and adversarial where the latter is generated by both Backdoor and

DolphinAttack algorithms. Next, each sample can be divided into mini-batches and the LID score can be computed for each mini-batch of these three groups with respect to their corresponding legitimate examples, by Eq. 3.17:

$$\mathrm{LID}(\mathbf{x}) = -\left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{r_i(\mathbf{x})}{r_k(\mathbf{x})} \right)^{-1} \tag{3.17}$$

where $\mathbf{x} \in \Re^{n \times m}$ is a 2D array, $r_i(\mathbf{x})$ refers to the distance between $\mathbf{x}$ and their nearest neighbors, $r_k(\mathbf{x})$ denotes the maximum of the neighbor distances, and k is the number of neighbouring samples. The LID scores of noisy and normal samples should be appended into negative class; and the LID scores of adversarial samples should be assigned to the positive class. Finally, a logistic regression can be trained on these two classes. The experiments carried out on 2D representations of audio signals in Section 3.5.1 show that the adversarial examples generated by Backdoor and DolphinAttack can change the true label, although they cannot be categorized as adversarial attacks because of two main reasons: (i) the adversarial examples lie in the subspace of legitimate and random noisy signals when they basically should lie into different sub-regions; (ii) since there is not an analytical or an optimization-based approach for computing small adversarial perturbations for high-dimensional audio, the values of such perturbations are actually generated manually or by greedy algorithms and therefore, this highly increases the chance of detecting the adversarial signal even by a simple defense model.

As it has been discussed so far, there are many open problems in crafting adversarial perturbations to raw audio signals and there is no reliable adversarial attack to 1D signals. This could also be interpreted as a good point if we disregard the fact that audio can be converted to a 2D representation (spectrogram) where strong adversarial attacks developed for images (e.g. FGSM, BIM, etc.) are quite applicable for 2D audio representations. This is a critical issue and poses a security concern for machine learning models for audio, either shallow (e.g. SVM) or deep learning models (e.g. CNNs). However, addressing the transferability of adversarial examples from 1D audio signals to 2D audio representations (or vice versa) is out of the scope of this paper. In fact, one of our goals in this paper is to assess the resiliency of machine learning

models based on different types of 2D audio representations to some strong adversarial attacks aiming to better understand their vulnerabilities.

3.3 2D Audio Representation

The vulnerability of machine learning models such as CNNs and long short-term memory networks on audio waveforms has been studied by Carlini and Wagner (Carlini & Wagner, 2018). They have shown the weaknesses of these models against FGSM-like adversarial attacks. However, the state-of-the-art for several audio tasks, such as music genre classification (Liu, Feng, Liu, Wang & Liu, 2019; Costa, Oliveira, Koerich, Gouyon & Martins, 2012), speaker identification (Sengupta, Yasmin & Ghosal, 2019), environmental sound classification (Sengupta *et al.*, 2019), etc. are based on 2D representation. This aroused our interest to evaluate the robustness of models based on 2D representations against adversarial attacks. To the best of our knowledge, the resiliency of 2D CNNs such as AlexNet and GoogLeNet, which have achieved the highest performances on environmental sound datasets, against adversarial attacks has not been studied in 2D representation spaces. To such an aim, we use Fourier and wavelet transforms to convert raw audio signals into 2D representations. The first transformation is used to produce short-time frequency spectrograms for training AlexNet and GoogLeNet (Boddapati *et al.*, 2017). We also use wavelet transform for producing more informative spectrograms, which after some pre-processing steps are used in the proposed approach to train an SVM classifier. A brief description of these two types of spectrogram is presented as follows.

Considering a discrete-time audio signal $a[n]$, where $n = 0, 1, \ldots, N-1$ denotes the number of samples and its decomposed signal S using Fourier (time-frequency) transform using $\left\{g_{\tau,\varrho}\right\}_{\tau,\varrho}$ atoms, as:

$$S[\tau, \varrho] = \left\langle a, g_{\tau,\varrho} \right\rangle = \sum_{n=0}^{N-1} a[n] g^*_{\tau,\varrho}[n] \tag{3.18}$$

where the operator $*$ denotes the complex conjugate, and τ, ϱ are time and frequency localization indices, respectively. This representation is widely used in sound and speech processing (Yu, Mallat & Bacry, 2008; Mallat, 2008). Given a Hanning window $H[n]$ of size ϑ which is shifted

by a step $u \leq \vartheta$, then $\left\{g_{\tau,\varrho}\right\}_{\tau,\varrho}$ in the latter equation can be defined as (Yu & Slotine, 2008):

$$g_{\tau,\varrho}[n] = H[n - \tau u] \exp\left(\frac{j2\pi\varrho n}{\vartheta}\right) \tag{3.19}$$

where $0 \leq \tau \leq N/u$ and $0 \leq \varrho \leq \vartheta$ denote bindings of time and frequency (scale) indices respectively. Finally, the Fourier spectrogram is represented as:

$$\mathbf{sp}_{\text{STFT}}[\tau, \varrho] = \log|S[\tau, \varrho]| \tag{3.20}$$

The final appearance of a spectrogram depends on the parameters τ and ϱ. Similar to this transform is the continuous wavelet transform (CWT) as denoted in Eq. 3.21:

$$CWT(\mho, z; a(t), \psi(t)) = \frac{1}{\sqrt{\mho}} \int_{-\infty}^{+\infty} a(t)\psi(\frac{t - z}{\mho})dt \tag{3.21}$$

where $\psi(t)$ denotes the mother wavelet and $\mho$, z, and t stand for scale, translation and time, respectively. The discretized representation of CWT is given by Eq. 3.22, and it is determined on a grid of $\mho$ scales and n discrete time with dilation parameter ρ.

$$DWT(\mho, n) = 2^{\mho/2} \sum_{\rho=0}^{n-1} a(\rho)\psi(2^{\mho}, \rho - n) \tag{3.22}$$

For ψ, we use Morlet function where $\mho$ is set to 0.8431:

$$\psi(t) = e^{-(\mho^2 t^2)/2} \cos(j\pi t) \tag{3.23}$$

Finally, the wavelet spectrogram can be obtained as:

$$\mathbf{sp}_{\text{DWT}}[\mho, z] = |DWT(\mho, z)|^2 \tag{3.24}$$

In summary, for an audio signal $a[n]$, there will be two different 2D representations: $\mathbf{sp}_{\text{STFT}}$ and $\mathbf{sp}_{\text{DWT}}$. Moreover, for the latter spectrogram we use three scales for the magnitude, which provide different visualization schemes: linear, logarithmic, and logarithmic real. Linear scale

highlights high-frequency magnitudes which represent high variation areas in the spectrogram. Logarithm scale highlights low-frequency information which expands distance of magnitudes in different scales. Finally, logarithm real scale highlights the energy of the signal which is related to the signal's mean.

3.4 A Robust Approach for 2D Audio Representation and Classification

In general, the current approaches for audio classification are able to achieve high accuracy but they are vulnerable to adversarial attacks, which means that they can be easily fooled by adversarial examples. Therefore, our aim is to design a novel approach for audio classification that provides a good trade-off between classification accuracy and low vulnerability to some of the most threatening adversarial attacks. The proposed approach for environmental sound classification has three main parts: spectrogram preprocessing, feature extraction, and classification. Fig. 3.2 presents an overview of the proposed preprocessing approach which, given an audio signal produces three spectrogram representations as output. The audio signal undergoes through color compensation, highboost filtering, dimensionality reduction, and smoothing and at the end, we have three enhanced spectrograms. Next, speeded up robust features (SURF) are extracted from zoning blocks that slide over the spectrograms as shown in Fig. 3.3. The geometrical distance of feature vectors is maximized by a K-means++ algorithm and finally a multiclass SVM trained on such features makes the prediction.

3.4.1 Spectrogram Preprocessing

The goal of the spectrogram preprocessing is threefold: (i) improve the accuracy of the front-end classifier; (ii) improve the robustness of the trained model against adversarial attacks; (iii) artificially increase the number of samples of the dataset. It starts by color compensation of the spectrogram $\mathbf{sp}_{\mathrm{DWT}}$ by mapping each spectrogram to three different color spaces: black-blue-green (BBG), purple-gold (PG), and white-black (WB) as shown in Fig. 3.4. Empirically, color compensation boosts and improves the final classification performance because it affects frequency coefficient values (i.e., pixel intensity in the spectrogram), though keeping their

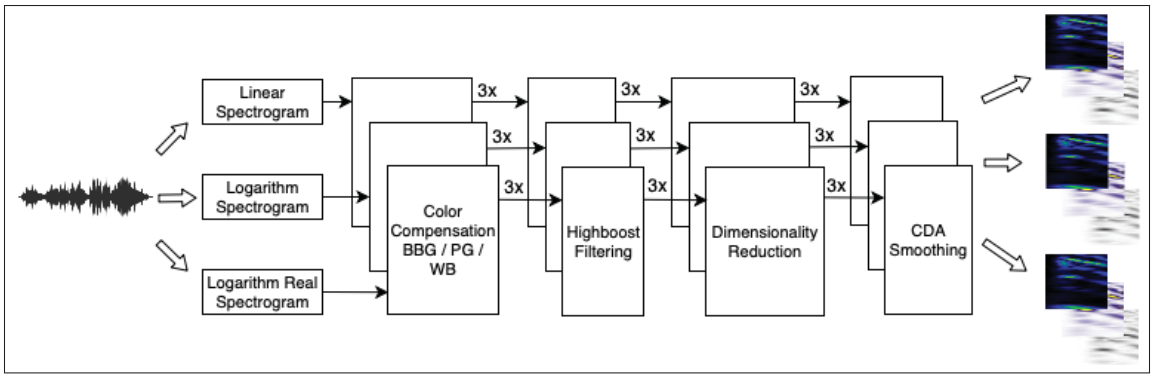

Figure 3.2 Overview of spectrogram generation and preprocessing. From a single audio waveform, three spectrogram representations are generated and processed through several blocks with the aim of enhancing the 2D representation.

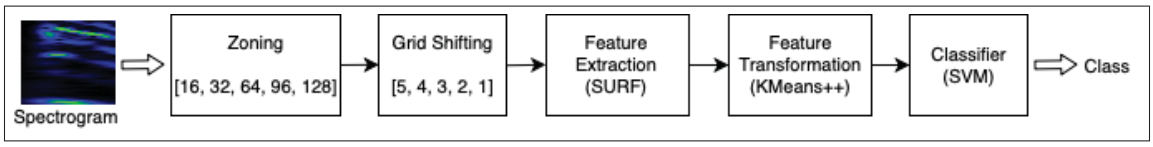

Figure 3.3 Overview of the proposed classification approach. Values in the first block indicate sizes of square zones (blocks) from 16×16 to 128×128. Stride values in the second block correspond to the zone sizes in the first block. For instance, a 96×96 block has stride 2, and so on.

distributions. The second preprocessing operation is highboost filtering (Gonzalez, 2016), which enhances color compensated spectrograms focusing on their high-frequency elements while maintaining low-frequency components. The output of the filter is denoted as $\mathbf{sp}_{\text{ENH}}$ which is given by Eq. 3.25.

$$\mathbf{sp}_{\text{ENH}} = (F_{ap} + cF_{hf}) \times \mathbf{sp}_{\text{DWT}} \tag{3.25}$$

where F_{hf} represents a high-pass filter (5×5 Laplacian operator) which is multiplied by a constant value c which acts as a scaling factor, and F_{ap} denotes an all-pass filter.

The three spectrogram representations and color compensations increase in nine times the number of samples into the datasets in addition to the pitch-shifting augmentation that is also applied, but on the 1D signal prior to the spectrogram representation. Pitch-shifting increases by eight times the number of samples. Therefore, to alleviate the computational

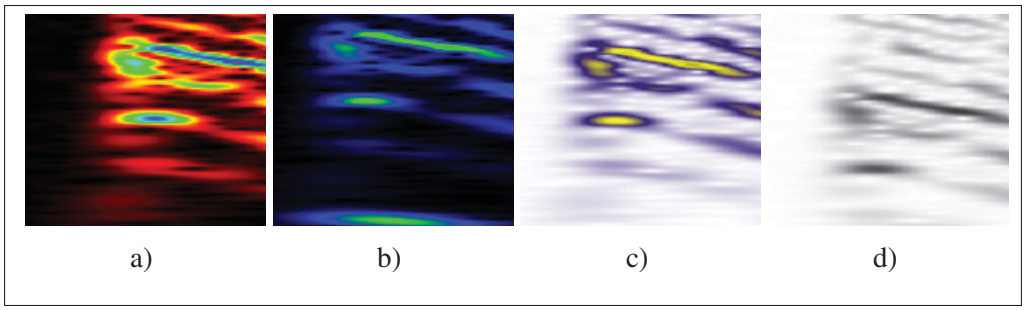

Figure 3.4 Spectrogram examples: (a) original; (b) black-blue-green (BBG); (c) purple-gold (PG); (d) white-black (WB)

complexity both in computing and storage, we reduce the dimensionality of the spectrograms. Though, there are many algorithms for such an aim, we use singular value decomposition (SVD) because of its pivotal properties in reducing the dimensionality of 2D data without changing the perceived visual appearance, if the reduction rank is chosen appropriately. Somewhat similar to the Fourier transform, SVD can describe a 2D matrix by basis functions in such a way that linear combination of these functions can reconstruct the original spectrogram (Esmaeilpour, Mansouri & Mahmoudi-Aznaveh, 2013). Basis functions in Fourier transform are sine and cosine, but SVD produces individual basis matrices for each given input. For an enhanced spectrogram $\mathbf{sp}_{\text{ENH}}$, SVD decomposes it as:

$$\mathbf{sp}_{\text{ENH}} = \sum_{i=1}^{m'} G_i U_i V_i^{\top} \tag{3.26}$$

where G, U, and V are derived matrices from decomposing $\mathbf{sp}_{\text{ENH}}$ into singular value, hanger, and aligner matrices, respectively. Also m' is the minimum dimension of the spectrogram either in width or height. The matrix G is a diagonal and its elements are in descending order which indicates the importance of hanger and aligner column vectors. The basis functions associated with $\mathbf{sp}_{\text{ENH}}$ are the product of U_i and $V_i^{\top}$ weighted by G_i. This allows us to reconstruct $\mathbf{sp}_{\text{ENH}}$ by its most important components, from low to high frequency components. By setting the m' in Eq. 3.26 to m'/n' where $n'>1$, we can make a balance between dimensionality reduction and quality of reconstruction. This operation actually acts as principal component analysis (Wall, Rechtsteiner & Rocha, 2003). Empirically, the magnitudes of G will be less than the

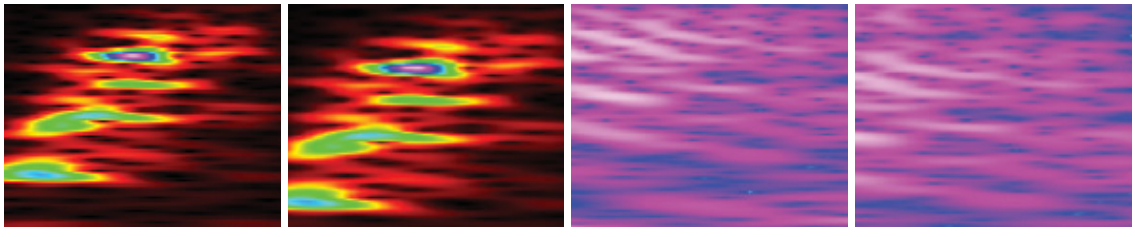

Figure 3.5 Dimension reduction effect: (a) linear magnitude representation; (b) reconstruction of (a) after reduction in half; (c) logarithmic magnitude representation; (d) reconstruction of (c) after reduction in half.

pixel precision (1/255 for an 8-bit representation) at indices around n'=2 and therefore they can be pruned without any visual impact on the spectrograms. Though this dimension reduction resizes spectrogram dimension to half, the quality of the reconstructed image is quite good, and differences are imperceptible to the human visual system (see Fig. 3.5). The outputs of the dimensionality reduction block in Fig. 3.2 are linear, logarithmic, and logarithmic real spectrograms visualized in three color spaces (BBG, PG, and WB) which are all reduced to half of their original dimension.

Though highboost filtering enhances high frequency components in spectrograms and therefore it leads to a better feature extraction, it may also boost noise, especially for the PG and WB color compensated representations. This problem can be minimized to some extent by the dimensionality reduction by SVD, but it is still necessary to improve the quality of the final compensated representations of spectrograms. For addressing this issue, highboost filtered spectrograms are smoothed using a denoising autoencoder with three convolution layers (Goodfellow, Bengio, Courville & Bengio, 2016). The main advantage of convolutional denoising autoencoder (CDA) over traditional smoothing algorithms is its flexibility in data adaptation and fine reconstruction. Besides, another important reason for using the CDA is to make spectrograms more robust against small adversarial perturbations which machine learning models are very sensitive to. The architecture of the proposed CDA depicted in Fig. 3.6 is data-dependent and it considers spectrograms of dimension 1167×765 as input. The architecture

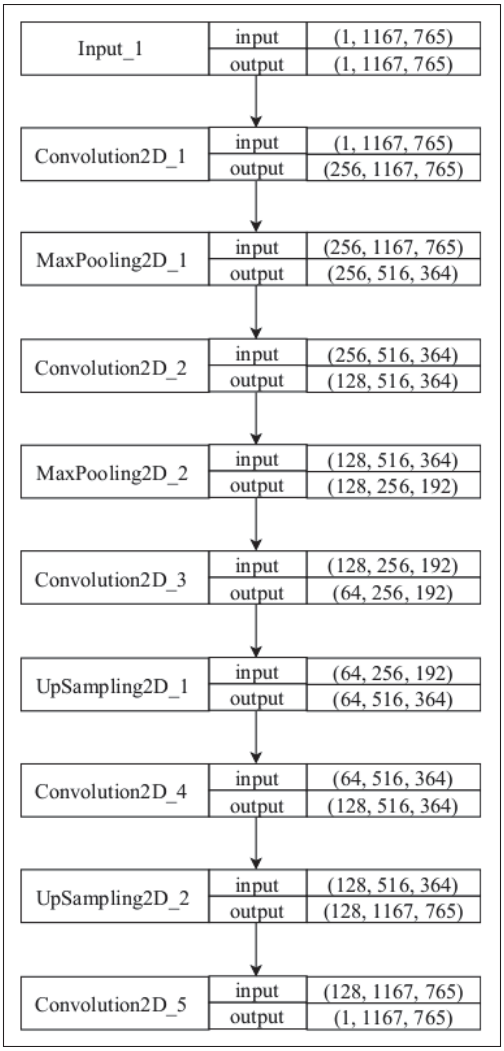

Figure 3.6 Architecture of our CDA

of the encoder shown in Fig. 3.6 has three convolutional layers with 5×5 receptive fields, stride 1, *Relu* activation function, dropout of 0.5, and two max pooling layers. For corrupting the input data, we used the spectrograms derived from SVD as well as the technique introduced by Vincent *et al.* (Vincent, Larochelle, Bengio & Manzagol, 2008).

Finally, after all these preprocessing steps, the enhanced spectrograms are ready to undergo to feature extraction and classification, as described in the following subsection. Besides that the enhanced spectrograms can also be used with pre-trained CNN architectures such as AlexNet or GoogLeNet, as described in Section 3.5.

3.4.2 Feature Extraction and Classification

The proposed approach includes five steps for feature extraction and classification as depicted in Fig. 3.3. The main idea is to extract features from a static sized moving aperture (a.k.a. grid shifting block) which spans a spectrogram with a dynamic stride within a block with dynamic size. Next, we maximize the geometrical distance among feature vectors of different classes and finally we train an SVM classifier on such an organized feature space. Since the proposed approach aims to achieving both classification accuracy and robustness against adversarial attacks, we have evaluated several handcrafted features and representation learning methods to finally come up with SURF instead of CNN features. Empirically, such a feature encoding outperforms DNN features (with/without convolution layers) both in terms of classification accuracy and robustness against adversarial attacks. Our main hypothesis relies on the nature of these features which are projected gradients compared to features generated by DNNs, which generally lead to high classification accuracy but empirically, they have a negative effect on the robustness of the trained model, which becomes quite vulnerable to adversarial attacks.

The first step is zoning, which divides a given spectrogram into zones that may vary from 16×16 to 128×128 pixels. Empirically, a zone size of 16×16 is small enough for capturing subtle pixel density changes and a zone size of 128×128 is preferable for regions with less high frequency components. Then, a sliding grid of size 8×8 will span through them. The stride of the sliding grid varies from five to one according to the zone size, with larger strides on larger zones. This scheme supports the idea of a detailed scanning of spectrograms aiming at extracting more discriminant features. Different values have been evaluated for the stride size and finally it ranges from one to five (see Fig. 3.7).

Different features could be extracted from each 8×8 grid. We also evaluated scale invariant feature transform (SIFT) as a feature extractor (Lowe, 1999) but decided to use SURF (Bay *et al.*, 2006) because it is much faster than SIFT in runtime, even if it provides fewer feature vectors compared to SIFT. We applied SURF on sliding grids within each zone as shown in Fig. 3.3, and at the end, each spectrogram zone is represented by a 64-dimensional feature

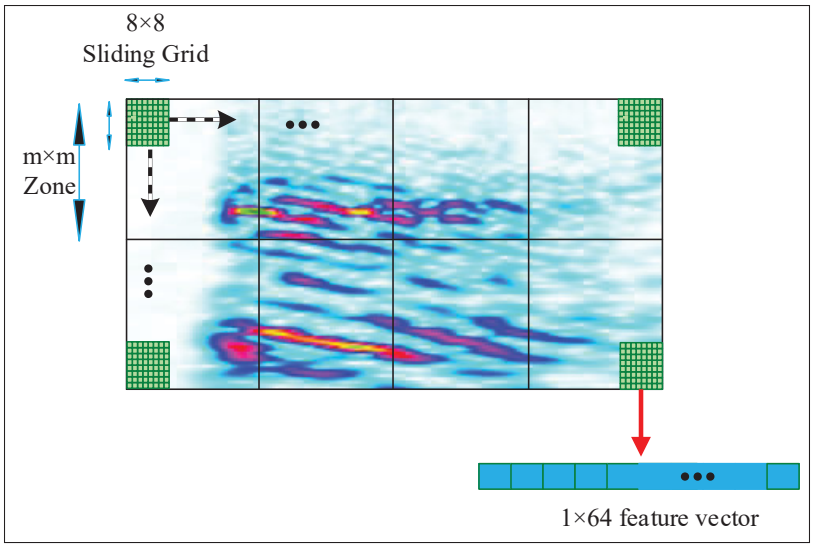

Figure 3.7 Example of a grid sliding over a spectrogram

vector. For increasing the inter-class geometrical distance among extracted feature vectors, the K-means++ algorithm (Arthur & Vassilvitskii, 2007) is used to cluster feature vectors into an organized distribution with respect to their geometrical linear distance. Once centroids are found by the clustering algorithm all feature vectors will be mapped into a distance space according to their centroids. We refer readers to (Coates & Ng, 2012) for further details. Finally, we train a multiclass SVM classifier with polynomial kernel on the transformed feature vectors. We have also evaluated the SVM with radial basis function (RBF) kernel which could not improve the accuracy. In the following section we evaluate the proposed approach on three datasets and compare the results with other state-of-the-art approaches.

3.5 Experimental Results

We have carried out several experiments on three benchmarking datasets with the aim of: (i) evaluating the detectability of the current adversarial attacks for 2D audio representations; (ii) assessing the performance of the proposed approach on the enhanced spectrograms and compare it with deep architectures (AlexNet and GoogLeNet) that have been used for audio classification; (iii) evaluating the resiliency of the proposed approach and the two deep architectures against

several types of adversarial attacks; (iv) characterizing the transferability of the adversarial audio attacks across two different classification paradigms, CNNs and SVMs.

The UrbanSound8K dataset has 8,732 audio samples of up to four seconds of 10 classes (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music). The ESC-50 dataset includes 2,000 5-second samples of 50 classes including major groups of animals, natural sound capes & water sounds, human non-speech sounds, domestic sounds, and exterior noises. The ESC-10 dataset is a subset of ESC-50 which includes 400 recordings of 10 classes (dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, and fire crackling).

3.5.1 Detectability of Adversarial Audio Attacks

The definition of an adversarial attack relies on whether the attack is easily identified or not. We have carried out some experiments to evaluate two of the most powerful adversarial attacks on audio: Backdoor and the DolphinAttack. For such an aim, we generated short-time Fourier transform (STFT), DWT, and cross recurrence plot (CRP) spectrograms for the audio samples of the UrbanSound8K dataset and computed the LID score considering different values of k as shown in Eq. 3.17. Basically, the LID score should be able to discriminate between negative and positive classes which means returning higher values. In other words, small values of the LID score denote an indistinguishable difference between positive and negative classes which can in turn be interpreted as positive classes may not be considered as adversarial. Table 3.1 shows that the differences between LID scores of positive and negative classes are quite small and it also shows the very low accuracy of the logistic regression classifier trained on these classes. As Table 3.1 shows, legitimate, noisy, and adversarial examples lie in the same subspace and in fact they lie in the same manifolds because they have very similar LID scores. In other words, the adversarial examples generated by both Backdoor and DolphinAttack are almost equivalent to examples corrupted by random noise, which basically does not seem to satisfy the definition of adversarial examples. Moreover, the performance of the logistic regression is quite low and

Table 3.1 LID score for different representations of UrbanSound8K samples. Mean difference is generated for two classes of negative (legitimate and random noisy) and positive (adversarial by Backdoor and DolphinAttack).

2D Representation type	k	Mean Difference of LID Scores	Classification Accuracy (AUC score %)
DWT	50	0.082	11.23
	75	0.071	10.04
	100	0.036	10.01
	125	0.032	09.46
STFT	50	0.076	13.05
	75	0.074	12.94
	100	0.066	12.92
	125	0.061	11.87
CRP	50	0.089	15.01
	75	0.084	14.56
	100	0.079	14.32
	125	0.078	13.77

shows poor discrimination between negative and positive classes which should be higher than 60%.

3.5.2 Accuracy and Resilience of CNNs and SVMs

Deep neural networks require a large amount of data for training. For increasing the size of datasets aiming at extracting more information from them, we augmented the number of samples by stretching (speeding up) and shrinking (slowing down) recordings in time (pitch shifting) using MUDA library (McFee *et al.*, 2015a). This is a common approach in sound processing which affects favourably the classifier's performance (Salamon & Bello, 2017). The scale values that were applied for pitch-shifting are: 0.5, 0.75, 0.9, 1.1, 1.25, 1.5, and 1.75. This operation increases the size of each dataset in eight times.

For generating the spectrogram $\mathbf{sp}_{\mathrm{STFT}}$, we used the approach suggested by Boddapati *et al.* (Boddapati *et al.*, 2017) by setting sampling frequency to 8 kHz, 16 kHz, and 8 kHz for ESC-10, ESC-50, and UrbanSound8K datasets, respectively. Also, the frame length was set to 50 ms (ESC-10), 30 ms (ESC-50), and 50 ms (UrbanSound8K) with a fixed overlapping of 50%.

Table 3.2 Scale operators (c) for color compensation

Benchmarking Dataset	Color Compensation	c (average)
ESC-10	BBG	0.57
	PG	0.74
	WB	0.46
ESC-50	BBG	0.81
	PG	0.79
	WB	0.58
UrbS8K	BBG	0.72
	PG	0.85
	WB	0.67

These values have been found after conducting exploratory experiments on these datasets. For generating the spectrogram $\mathbf{sp}_{\mathrm{DWT}}$, we used 256 frequency bins with a Morlet mother function as proposed by Cowling and Sitte (Cowling & Sitte, 2003) and linear, logarithmic, and logarithmic real magnitude scales for enhancing high, low and medium frequencies, respectively. The scale operators c as described in Eq. 3.25, are shown in Table 3.2. The SVM uses a quadratic kernel with the cost parameter $\|c\| \leq 0.1$ and the kernel parameter $\|\gamma\| < 0.003$. Besides the quadratic kernel, we also evaluated a linear SVM, which is referred simply as SVM in several tables in this section. We have used the scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg et al., 2011b) package for implementing SVMs.

In the first experiment, we trained AlexNet and GoogLeNet with the same setup proposed by Boddapati *et al.* (Boddapati *et al.*, 2017) which leads to the highest classification performance reported in the literature for 2D representations. These two deep convolutional neural networks were trained on a linear pooling of STFT ($\mathbf{sp}_{\mathrm{STFT}}$), MFCC ($\mathbf{sp}_{\mathrm{MFCC}}$), and CRP ($\mathbf{sp}_{\mathrm{CRP}}$) spectrograms, as:

$$\mathbf{sp}_{\mathrm{POOL}} = \mathrm{clip}\left(\mathbf{sp}_{\mathrm{STFT}} + \mathbf{sp}_{\mathrm{MFCC}} + \mathbf{sp}_{\mathrm{CRP}}, [0, 1]\right) \tag{3.27}$$

where $\mathbf{sp}_{\mathrm{POOL}}$ denotes the resulting pooled spectrogram which values outside the range [0,1] are clipped to the value at the boundary of the range. These spectrograms are computed for the three environmental sound datasets (ESC-10, ESC-50, and UrbanSound8K) after the augmentation procedure. In addition to training our classifier on the pooled representation, referred to as

Table 3.3 Mean classification accuracy (5-fold CV) of four classifiers on two representation spaces: POOL and DWT. The best performances are shown in bold.

Benchmarking Dataset	2D Representation type	Mean Accuracy (%)			
		GoogLeNet	AlexNet	SVM	Proposed
ESC-10	POOL	**83.19**	82.54	64.23	78.31
	DWT	**83.21**	82.90	70.45	79.10
ESC-50	POOL	**71.36**	64.09	52.37	60.10
	DWT	**71.20**	66.41	55.09	60.41
UrbS8K	POOL	**91.08**	90.06	72.03	86.15
	DWT	86.85	**90.10**	72.89	86.39

POOL, we also trained it on the enhanced 2D representation space as shown in Fig. 3.2, referred to as DWT. These two representations are also evaluated for AlexNet and GoogLeNet. In other words, we evaluate the performance of AlexNet and GoogLeNet on the spectrograms obtained from our data preprocessing approach. These two experiments are executed using 5-fold cross validation with a ratio of 0.2 for testing. We used four parallel GPUs GTX580 based on an implementation based on (Krizhevsky *et al.*, 2012). We stopped training after 83 epochs using early stopping for AlexNet and GoogLeNet. The results achieved by these two classifiers are reported in Table 3.3. As Table 3.3 shows, AlexNet and GoogLeNet have achieved the best performances for both representation spaces, although the proposed approach presents competitive results. The differences between the best deep model and the proposed approach range from 4.32% for UrbanSound8K to 11.02% for ESC-50. We also repeated this experiment with 10-fold cross validation as suggested in (Salamon & Bello, 2017), but the results were very close to those reported in Table 3.3.

However, a high accuracy does not translate to a high robustness against adversarial attacks. In Table 3.4, we assess the robustness of the classifiers of Table 3.3 against several adversarial attacks as well as the transferability of such adversarial attacks across different models. For such an aim we have developed the FGSM, BIM-a, BIM-b, and CWA adversarial attacks (deep model attacks) for AlexNet and GoogLeNet and the EA and regular Evasion attacks (SVM attacks) for SVM classifiers. The total number of adversarial examples crafted using each attack for different datasets is equivalent to the number of samples in the legitimate dataset. In other words, for

each legitimate sample, one adversarial example is crafted by each adversarial attack algorithm. Since FGSM and CWA are targeted, adversarial examples of these two attacks are crafted toward a random wrong label. This not only makes our evaluations fair against non-targeted attacks, but also reduces the cost of crafting adversarial examples of datasets with more than 10 classes, which is the case of the ESC-50 dataset that has 50 classes. Then, these crafted examples are fed to both deep learning and SVM models to compute the ratio of successful fooling over the total number of adversarial examples (fooling rate) in a black-box scenario.

Table 3.4 also shows the transferability of adversarial attacks crafted to attack deep models to attack SVM models and vice-versa. A high adversarial transferability rate represents a serious threat for data-driven classifiers. In other words, a reliable classifier should not only be robust against adversarial attacks designed to fool its own type of model, but it should also be reasonably resistant against attacks designed to attack other types of model. Table 3.4 shows the results achieved on both experiments. The mean fooling rate, which measures the success rate of adversarial examples in fooling the machine learning models in terms of the percentage of adversarial samples misclassified by the models is computed for comparing the performance of CNNs and SVMs against the six adversarial attacks. Table 3.4 shows that both for both CNNs and SVMs are quite vulnerable to the adversarial attacks designed to attack its own model, with fooling rates higher than 90%. However, the proposed approach not only is quite robust, but also has the lowest fooling rate against adversarial attacks (EA and LFA) designed for such a model, with fooling rates between 58.15% and 71.64%. Table 3.4 also reveals that there is a higher chance of fooling SVM models by deep attacks compared to fooling AlexNet and GoogLeNet by adversarial examples crafted by EA or LFA. Additionally, AlexNet is more robust against SVM-based adversarial attacks compared to GoogLeNet, though its mean accuracy is a little lower than GoogLeNet.

Table 3.5 shows average rankings of our evaluation metrics of recognition accuracy and fooling rate with respect to the statistics provided in Table 3.4. Regarding this table, the smaller the $\bar{r}$ is, the better are the accuracy and the fooling rates. Although the proposed approach appears in third place in the mean accuracy rank, it is the first one in resiliency against the six types of adversarial

Table 3.4 Mean fooling rate (5-fold CV) of two CNNs and two SVMs against six strong adversarial attacks. The best performances are shown in bold (lowest values).

Dataset (Representation)	Adversarial Attack	Mean Fooling Rate (%)			
		GoogLeNet	AlexNet	SVM	Proposed
ESC-10 (POOL)	FGSM	95.23	94.04	60.78	**43.12**
	BIM-a	94.07	90.13	61.68	**48.60**
	BIM-b	94.26	91.30	62.46	**46.03**
	CWA	95.89	93.66	94.01	**51.77**
	LFA	**51.23**	63.01	94.43	60.47
	EA	**43.79**	44.12	94.14	58.34
ESC-10 (DWT)	FGSM	94.30	93.36	64.05	**50.02**
	BIM-a	92.15	92.87	59.57	**51.13**
	BIM-b	93.58	92.33	57.92	**43.07**
	CWA	95.36	94.89	64.35	**53.18**
	LFA	57.36	**56.35**	95.58	71.64
	EA	49.66	**48.00**	92.89	61.78
ESC-50 (POOL)	FGSM	96.78	95.61	69.22	**51.99**
	BIM-a	95.01	96.08	67.17	**50.20**
	BIM-b	94.77	95.17	69.71	**50.03**
	CWA	96.02	97.14	72.10	**53.04**
	LFA	62.12	66.35	95.27	**60.25**
	EA	55.47	**52.01**	95.94	59.03
ESC-50 (DWT)	FGSM	96.30	95.80	66.16	**50.01**
	BIM-a	93.36	94.05	69.02	**49.36**
	BIM-b	91.25	92.53	67.11	**45.92**
	CWA	95.73	94.11	70.09	**49.31**
	LFA	60.08	**58.01**	92.21	62.84
	EA	51.37	**49.61**	90.36	58.15
UrbS8K (POOL)	FGSM	94.68	93.22	60.50	**45.17**
	BIM-a	94.65	95.32	58.22	**42.36**
	BIM-b	90.22	91.24	53.39	**42.16**
	CWA	92.08	93.62	**60.17**	60.25
	LFA	**55.01**	78.36	96.14	65.35
	EA	44.02	**41.07**	95.16	62.30
UrbS8K (DWT)	FGSM	94.14	93.02	57.31	**48.33**
	BIM-a	92.43	93.21	62.01	**51.07**
	BIM-b	94.01	93.61	63.32	**53.03**
	CWA	95.27	93.84	62.14	**50.48**
	LFA	**54.33**	55.03	92.06	63.52
	EA	47.01	**45.50**	91.02	59.01

attacks. Therefore, this indicates a good trade-off between accuracy and resiliency. This is also shown in Fig. 3.8, where the proposed approach is the one closest to the origin (zero error rate

Table 3.5 Average ranking considering the mean accuracy and the fooling rate for all models, datasets and adversarial attacks

Classification Approach	Mean Accuracy		Fooling Rate	
	$\bar{r}$	Sorted Rank	$\bar{r}$	Sorted Rank
GoogLeNet	1.17	1	2.97	4
AlexNet	1.83	2	2.78	3
SVM	4.00	4	2.67	2
Proposed	3.00	3	1.61	1

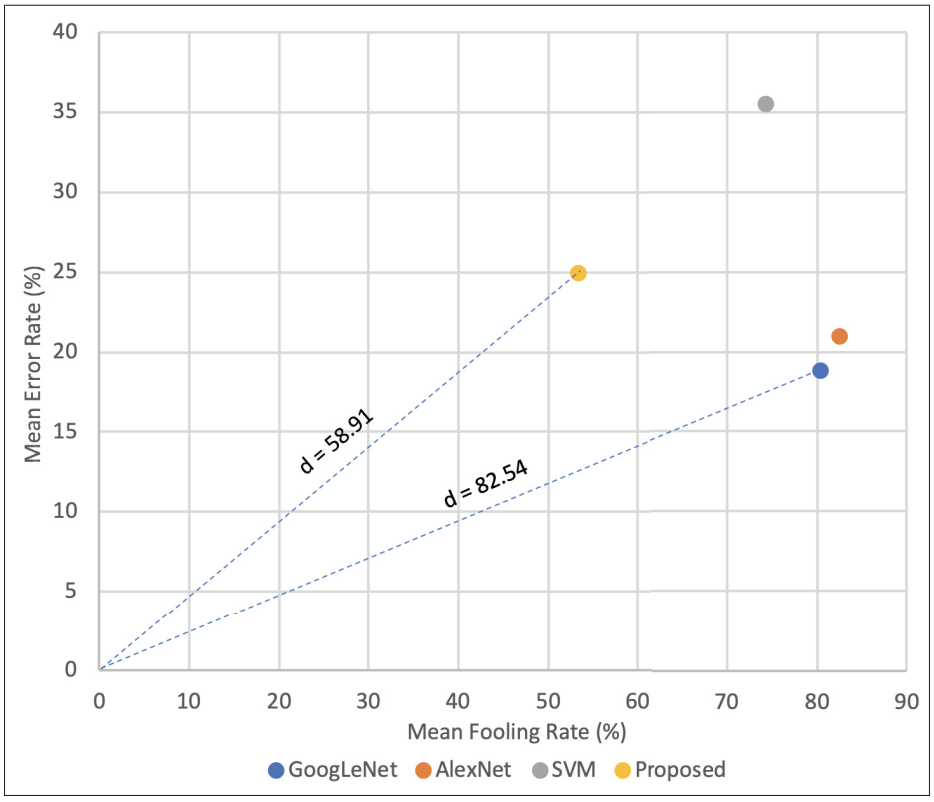

Figure 3.8 Model Comparison over the representations of Table 3.3

and zero fooling rate) according to the Euclidean distance (d = 58.91). Fig. 3.8 also shows that while the mean error rate of the proposed approach is 6.07% higher than GoogLeNet, the proposed approach is 26.96% more robust to adversarial attacks than GoogLeNet. Furthermore, the mean error rate of the proposed approach is 10.57% lower than the SVM and it is also

Table 3.6 The average effect of removing each module from Fig. 3.2 on the mean accuracy and robustness of the proposed model against deep and SVM adversarial attacks. Positive (+) and negative (−) effects are shown by their signs.

Module	Mean Accuracy (%)	Robustness Against (%)	
		SVM Adversarial Attacks	Deep Adversarial Attacks
Spectrogram Visualization	−16.47	−4.07	−3.14
Color Compensation	−7.36	−0.36	+2.64
Highboost Filtering	−9.52	−0.75	−1.96
SVD	−8.21	−6.18	−4.17
CDA	−7.94	−9.18	−6.33

20.98% more robust to adversarial attacks. Notwithstanding the good trade-off achieved by the proposed approach, there is still a large room for improvements.

3.5.3 Analysis of the Proposed Approach

The proposed approach provides the best trade-off between accuracy and resilience to adversarial attacks than deep models and SVM according to the proposed metric shown in Fig. 3.8. For understanding the reason(s) of such a best trade-off, we dig into the preprocessing (Fig. 3.2) of the proposed approach. We safely remove each module (or submodule) from the preprocessing part and measure its positive or negative contribution to the mean accuracy and robustness against the six types of adversarial attacks. Table 3.6 reveals that the proposed approach takes advantage of both CDA and SVD compression. The most straightforward impact of these two operations is in affecting (smoothing) high frequency components where subtle changes of adversarial examples probably lie on. It has been proved that autoencoders can clean adversarial examples and therefore defend the targeted trained models from the adversarial attacks (Nayebi & Ganguli, 2017; Meng & Chen, 2017). Moreover, for measuring the effect of the first two modules of Fig. 3.3 on final classification performance, we carried out some additional experiments including removing them and changing block size and grid shifting stride on a 5-fold cross validation. In Table 3.7, we only report some of the highest mean accuracy with respect to zoning size and grid shifting stride.

Table 3.7 The effect of selected zoning size and shifting grid length on the overall recognition accuracy of the proposed approach on DWT representation of the UrbanSound8K dataset

Zoning static Size	Sliding Grid Stride (ordered)	Mean Accuracy (%)
[16, 32, 64, 96, 128]	[1, 2, 3, 4, 5]	**79.33**
[16, 32, 64, 96, 128]	[2, 2, 2, 2, 2]	77.29
[8, 16, 32, 64, 128]	[1, 2, 3, 4, 5]	76.18
[16, 32, 64, 96, 128]	[4, 3, 3, 3, 4]	74.22
[32, 64, 128]	[3, 2, 1]	73.91
[64, 96, 128]	[3, 2, 1]	72.63
None	None	70.92

3.6 Conclusion

In this paper, we discussed the serious threat that adversarial attacks may pose to machine learning models trained either on 1D or 2D audio representations. While there is no reliable adversarial attack on raw audio signals, there is a bijective relation between 1D signals and spectrograms which opens the avenue for adversarial transferability between these two representation spaces and that poses a real security concern. Besides that considering that the majority of state-of-the-art approaches for audio classification rely on 2D representations, most of them based on CNNs originally designed for image classification tasks, we showed that CNNs trained on spectrograms of environmental sound signals achieve state-of-the-art performance in terms of accuracy. However, these CNNs are not reliable at all, as they can be easily fooled by adversarial examples, with fooling rates higher than 90%.

Therefore, we proposed a novel approach for environmental sound classification based on 2D representations that provides a good tradeoff between accuracy and resiliency to the most powerful adversarial attacks designed to fool both deep neural models and SVMs. The proposed approach was compared to AlexNet, GoogLeNet, and a linear SVM classifier on three publicly available datasets. The highest mean recognition rates were achieved by GoogLeNet (81.15%), AlexNet (79.15%), the proposed approach (75.08%), and the linear SVM (64.51%), respectively. However, in addition to the competitive accuracy, the proposed approach outperforms by far all three mentioned classifiers in terms of robustness against adversarial attacks since the mean

fooling rates for these four models are 95.15%, 94.36%, 50.56%, and 66.74% considering deep attacks and 52.62%, 54.79%, 61.89%, and 93.77% considering SVM attacks. However, as shown in Fig. 3.8, there is still a large room for improvements. As a future study, we are interested in employing reactive adversarial detection algorithms (e.g., LID detector) as a postprocessing operation aiming at increasing the robustness of the proposed approach.

We are also inclined to explore the resiliency of our classification scheme for raw audio signals rather than spectrograms against audio attacks and measure its capability against audio played back over the air. To this end, we may need to remove/add some preprocessing steps which have shown positive impacts on the robustness of the proposed approach against adversarial attacks (e.g. CDA); and consequently, simplify our approach which requires several steps of processing. Another important aspect that deserves further studies is the adversarial example transferability bijectively from 1D audio signal to 2D spectrograms and vice versa. In other words, we would like to explore the possibility of crafting adversarial audio examples for a model trained on 1D signals and transfer such an attack to the 2D representation to be able to fool a 2D model trained on spectrograms, and also the other way around. Since many audio classification approaches implement different types (ensemble) of data-driven models (both 1D and 2D) aiming at improving their prediction confidence, hence if a crafted adversarial example can fool both 1D and 2D models, it may constitute a true threat to several sound recognition/processing systems and devices (e.g. voice id devices).

CHAPTER 4

FROM SOUND REPRESENTATIONS TO MODEL ROBUSTNESS: A COMPREHENSIVE DISCUSSION

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper Submitted for Publication to « Elsevier Applied Acoustics » in March 2021.

Abstract

This paper investigates the impact of different standard environmental sound representations (spectrograms) on the recognition performance and adversarial attack robustness of a victim residual convolutional neural network. Averaged over various experiments on three comprehensive environmental sound datasets, we found the ResNet-18 model outperforms other deep learning architectures such as GoogLeNet, AlexNet, and variants of residual and recurrent configurations both in terms of classification accuracy and the number of training parameters. Therefore, we opted for this advanced model as our front-end classifier for subsequent investigations. Herein, we measure the impact of different settings required for generating more informative Mel-frequency cepstral coefficient (MFCC), short-time Fourier transform (STFT), and discrete wavelet transform (DWT) representations on our front-end model. This measurement involves comparing the classification performance over the adversarial robustness. We demonstrate an inverse relationship between recognition accuracy and model robustness against six benchmarking attack algorithms on the balance of average budgets allocated by the adversary and the attack cost. Moreover, our experimental results show that while the ResNet-18 model trained on DWT spectrograms achieves the highest recognition accuracy, attacking this model is relatively more costly for the adversary than other 2D representations.

4.1 Introduction

Developing reliable sound recognition algorithms for real-life applications has always been a significant challenge for the signal processing community (Marchegiani & Posner, 2017; Salamon, MacConnell, Cartwright, Li & Bello, 2017; Radhakrishnan *et al.*, 2005). For analyzing the surrounding scene either for surveillance (Valenzise *et al.*, 2007) or multimedia sensor networks (Steele *et al.*, 2013), there is a constant need to understand environmental events. Raised by these concerns, several unsupervised (Salamon & Bello, 2015b) and supervised (Salamon & Bello, 2017) algorithms have been devised for classifying environmental sounds. During the last decades, there has been increasing attention toward developing automatic environmental sound classifiers. Presumably, this is due to its vast applications in smart acoustic sensor network development (Mydlarz *et al.*, 2017), surveillance scene monitoring (Radhakrishnan *et al.*, 2005; Cristani *et al.*, 2004), IoT-based noise reduction (Shah *et al.*, 2019), smart city safety (Ciaburro & Iannace, 2020; Shah *et al.*, 2018), and context-aware computing (Chandrakala *et al.*, 2021; Chu *et al.*, 2009a; Toffa & Mignotte, 2020). Towards developing reliable classification algorithms for such tasks, the impact of adversarial attacks on the deep learning (DL) classifiers trained on environmental sounds should be investigated. In other words, developing reliable environmental sound classifiers requires the study of adversarial attacks in greater detail to account for the impact of such attacks on different sound representations. This is our main motivation for setting the framework of this paper to environmental sounds encompassing a broad spectrum of urban sounds.

With the proliferation of DL algorithms during the last decade for image-related tasks, many publications on 2D audio representations (spectrograms) have been released. The DL architectures primarily developed for computer vision applications have been well adapted for sound recognition tasks with recognition accuracy competitive to human understanding. However, such algorithms require large amounts of training data. In response, many low-level data augmentation approaches have been introduced to allow an appropriate training of DL models and improve their performance on sound-related tasks (Salamon & Bello, 2017). These approaches apply directly to audio waveforms affecting low-level sampled data points of the audio signal, which may

not necessarily improve the performance of the front-end classification models (Esmaeilpour, Cardinal & Koerich, 2020a). High-level data augmentation approaches have been developed to tackle this problem, which are particularly useful for 2D audio representations (Kaneko, Takaki, Kameoka & Yamagishi, 2017; Mathur, Isopoussu, Kawsar, Berthouze & Lane, 2019). Experimental results on a variety of environmental sound datasets attest considerable positive impact of high-level data augmentation on overall performance of DL classifiers (e.g., AlexNet (Krizhevsky *et al.*, 2012), GoogLeNet (Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke & Rabinovich, 2015), etc.) (Esmaeilpour *et al.*, 2020a).

Unfortunately, recent studies have demonstrated the vulnerability of these convolutional neural networks (ConvNets) trained on 2D representations of audio signals against adversarial attacks (Esmaeilpour *et al.*, 2020). They have shown that crafted adversarial examples are transferable among dense ConvNets and support vector machines (SVM). That poses potential harm for sound recognition systems, especially when the highest recognition accuracy has been reported on 2D representations over raw 1D audio signals (Boddapati *et al.*, 2017). This threat negatively affects the reliability of DL models designed for applications based on sound classification, particularly IoT-related tasks in an environmental setting (Zamil, Samarah, Rawashdeh, Karime & Hossain, 2019).

Toward proposing reliable classifiers, there have been some debates and case studies on the link between intrusion of adversarial examples and loss functions for some victim classifiers (Carlini & Wagner, 2017b). It has been shown that the integration of more convex loss functions in the victim model (or in the surrogate counterpart) might increase the chance of crafting more potent adversarial examples (Carlini & Wagner, 2017b). However, it might also depend on some other key factors such as the properties of the classifier, input sample distribution, adversarial setups, etc. To study other potential links, we evaluate the robustness and the transferability of some state-of-the-art ConvNets against adversarial attacks trained on different 2D environmental sound representations. Our primary front-end ConvNet is ResNet-18 architecture because of its superior recognition performance compared to other ConvNet architectures. We discuss this in

Section 4.4.2 and briefly report our findings on different dense architectures such as GoogLeNet and AlexNet in Section 4.5.

The main novelty in this paper is investigating classifier response to different representations both in terms of recognition accuracy and robustness against adversarial attacks. This helps to yield more reliable classifiers without running any costly adversarial defense algorithm. More specifically, we make the following contributions:

1. we show that models achieve higher recognition accuracy on the DWT representation compared to STFT and MFCC averaged over different spectrogram settings for three comprehensive environmental sound datasets;
2. we identify major spectrogram settings which considerably affect the cost of attack (the number of required gradient computations) averaged over budgets;
3. we characterize the existence of an inverse relation between recognition accuracy and robustness of the victim models against six strong targeted and non-targeted benchmarking adversarial attacks. On average, models with higher recognition accuracies undergo higher fooling rates;
4. we demonstrate that compared to DWT and STFT, the MFCC has a relatively lower adversarial transferability ratio among three advanced DL architectures.

The rest of the paper is organized as follows. In Section 4.2, we briefly review some strong adversarial attacks for 2D audio representations. Then, explanations on different 2D audio representations that have been used in the experiments are summarized in Section 4.3. Next, experimental results and associated discussions are presented in Section 4.4. Finally, the conclusions and perspectives of future work are presented in the last section.

4.2 Adversarial Attacks

Assuming we have a sound recognition system that employs a classifier trained on legitimate spectrograms. In the following, we explain how crafted adversarial spectrograms can pose security concerns for this system.

1. *White-box scenario:* The adversary has full access to the entire system details, including audio dataset, classifier architecture, potential tuning parameters, required hyperparameters, and complete weight vectors. Therefore, the adversary can easily feed adversarial spectrograms to the model and fool it toward any incorrect target label.
2. *Black-box scenario:* The adversary does not have access to the system mentioned above details. Thus, the adversary can only input a 1D signal to the system and receive a predicted label. In this scenario, the adversary can reconstruct an audio signal from an adversarial spectrogram (with or without a surrogate model) and feed it to the system. Since the model is trained on spectrograms, the system first converts the input audio into a spectrogram that embeds the adversarial perturbation. This reconstruction does not pose a technical difficulty since spectrogram and 1D signal are dual, and there are plenty of straightforward approaches for reconstructing one from another. However, this spectrogram can also fool the model toward any wrong label defined by the adversary (see a relevant study by Koerich *et al.* (2020)).

4.2.1 Adversarial Attack For Environmental Sound Classifiers

In practice, adversarial attacks exist both for 1D signals (Li, Wu, Liu, Chen & Yuan, 2020b) and their associated 2D representations (Esmaeilpour *et al.*, 2020). This paper focuses on the latter since from decades ago, spectrograms (generated from MFCC, STFT, DWT) have been standard representations for different audio and speech processing tasks, particularly classification. Besides, spectrogram and 1D signal are duals (bijectively convertible), and the highest recognition accuracy on the benchmarking environmental sound datasets have been reported for models trained on the 2D representations (Boddapati *et al.*, 2017). Finally, since spectrograms are RGB matrices similar to natural images and the adversarial attacks developed in the computer vision domain are applicable on spectrograms.

Technically, an adversarial attack can be formulated as an optimization problem toward achieving a minimal perturbation parameter δ as stated in Eq. 4.1 (Szegedy *et al.*, 2014).

$$\min_{\delta} \quad f^*(\mathbf{x}+\delta) \neq f^*(\mathbf{x}) \tag{4.1}$$

where $\mathbf{x}$ and f^* denote a legitimate random spectrogram and the post-activation function of the victim classifier, respectively. The optimal value for δ should be as small as possible to not being perceivable by humans, although distinguishing the applied perturbation on 2D audio representations such as spectrograms is complicated. Many attack algorithms that satisfy such an imperceptibility constraint have been proposed in white and black-box scenarios. In this paper, we briefly go over six strong targeted and non-targeted adversarial attacks, which are well adapted to sound recognition models trained on 2D audio representations (Esmaeilpour *et al.*, 2020). We use the average fooling rate of these attacks, a standard metric for assessing the robustness of victim ConvNets trained on different 2D audio representations.

4.2.2 Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)

Szegedy *et al.* (Szegedy *et al.*, 2014) argue that the viability of fooling deep neural networks with fake examples is due to their extremely low probability because such examples are rarely seen in a given dataset. That could be understood as the pitfall of deep networks in low generalizability to unseen but very similar samples. However, they propose an optimization algorithm to mislead finely trained DL models, based on Eq. 4.2:

$$\min_{\mathbf{x}'} c\,\|\delta\|_2 + J_{\mathbf{w}}(\mathbf{x}', l') \tag{4.2}$$

where c is a positive scaling factor achievable by the line search strategy, $\mathbf{x}'$ denotes the associated crafted adversarial example, l' refers to its target label, and $J_{\mathbf{w}}$ denotes the loss function for updating weights ($\mathbf{w}$). There are various choices for this function, such as cross-entropy loss or any other surrogate function. The solution to this optimization problem is quite costly, and it has been proposed to use the L-BFGS optimizer, subject to $0 \leq \mathbf{x}' \leq M$ where M refers to the

maximum possible intensity in a spectrogram. This attack is the baseline for the adversarial algorithms that are subsequently presented.

4.2.3 Fast Gradient Sign Method (FGSM)

Goodfellow *et al.* (Goodfellow *et al.*, 2015) explain the existence of adversarial examples with linear nature of deep neural networks, even those with super-dense hidden layers. Toward this claim, they proposed a fast optimization algorithm based on Eq. 4.3:

$$\mathbf{x}' \leftarrow \mathbf{x} + \delta \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}, l)) \tag{4.3}$$

where δ is a small constant for controlling the applied perturbation to the legitimate sample $\mathbf{x}$. Different choices of ℓ_p norms can be integrated into the FGSM attack, and the adversary should make a trade-off between high similarities and a large enough perturbation to be able to fool a model. The formulation of Eq. 4.3 for ℓ_2 norm is shown in Eq. 4.4.

$$\mathbf{x}' \leftarrow \mathbf{x} + \delta \frac{\nabla_{\mathbf{x}} J(\mathbf{x}, l)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}, l)\|} \tag{4.4}$$

where for satisfying the constraint $\mathbf{x}' \in [0, M]$, the resulting adversarial spectrogram should be clipped or truncated. This white-box adversarial attack is targeted toward a pre-defined wrong label by the adversary in a one-shot scenario.

4.2.4 Basic Iterative Method (BIM)

This non-targeted adversarial attack (Kurakin *et al.*, 2016) is, in fact, the iterative version of the FGSM optimization algorithm, which crafts and positions potential adversarial examples ideally outside of legitimate subspaces via optimizing Eq. 4.5 for δ:

$$\mathbf{x}'_{n+1} \leftarrow \text{clip}_{\mathbf{x},\delta} \left\{ \mathbf{x}'_n + \delta \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}_n, l)) \right\} \tag{4.5}$$

where clip is a function for keeping generated examples within the range $[\mathbf{x}-\delta, \mathbf{x}+\delta]$ as defined in Eq. 4.6.

$$\min\left\{M, \mathbf{x}+\delta, \max\{0, \mathbf{x}-\delta, \mathbf{x}'\}\right\} \tag{4.6}$$

where M=255 for 8-bit RGB visualization of spectrograms.

There are two implementations for this optimization algorithm either by optimizing up to reach the first adversarial example (BIM-a) or continuing optimizing to a predefined number of iterations (BIM-b). The latter usually generates stronger adversarial examples, though it is more costly since it usually requires more callbacks. Both BIM attacks are iterative and white-box algorithms minimizing Eq. 4.5 for optimal perturbation δ measured by ℓ_∞ norm.

4.2.5 Jacobian-based Saliency Map Attack (JSMA)

Similar to the FGSM attack, this algorithm also uses gradient information for perturbing the input taking advantage of a greedy approach (Papernot *et al.*, 2016d). This attack is targeted toward a pre-defined wrong label (l'). In fact, it optimizes for $\arg\min_{\delta_\mathbf{x}} \|\delta_\mathbf{x}\|$ subject to $f^*(\mathbf{x}+\delta_\mathbf{x}) = l'$ (optimizing with ℓ_0). There are three steps in developing JSMA adversarial examples. First, computing the derivative of the victim model as Eq. 4.7.

$$\nabla f(\mathbf{x}) = \frac{\partial f_j(\mathbf{x})}{\partial x_i} \tag{4.7}$$

where x_i denotes pixels intensities. Second, a saliency map should be computed to detect the least effective pixel values for perturbation according to the desired outputs of the model. Specifically, the saliency map for pixels in cases where $\partial f_l(\mathbf{x})/\partial \mathbf{x}_i < 0$ or $\sum_{j\neq l} \partial f_j(\mathbf{x})/\partial \mathbf{x}_i > 0$ should be set to zero since there are detectable variations, otherwise:

$$S_{map}(\mathbf{x}, l')[i] = \frac{\partial f_l(\mathbf{x})}{\partial \mathbf{x}_i}\left|\sum_{j\neq l'} \frac{\partial f_j(\mathbf{x})}{\partial \mathbf{x}_i}\right| \tag{4.8}$$

where S_{map} denotes the saliency map for every given spectrogram $\mathbf{x}_i$ and target label l'_i. The last step of the JSMA is applying the perturbation on the original input according to the achieved map.

4.2.6 Carlini and Wagner Attack (CWA)

This is an iterative and white-box adversarial algorithm (Carlini & Wagner, 2017b), which can use three types of distance metrics: ℓ_0, ℓ_∞, and ℓ_2 norms. This paper focuses on the latter distance measure making the algorithm very strong even against the distillation network. The optimization problem in this attack is given by Eq. 4.9.

$$\min_{\delta} \|\mathbf{x}' - \mathbf{x}\|_2^2 + c f(\mathbf{x}') \tag{4.9}$$

where c is a constant value as explained in Eq. 4.2. Assuming the target class is l' and $G(\mathbf{x}')_i$ denotes the logits of the trained model f before softmax activation corresponding to the i-th class, then:

$$f(\mathbf{x}') = \max \left\{ \max_{i \neq l'} \{G(\mathbf{x}')_i\} - G(\mathbf{x}')_{l'}, -\kappa \right\} \tag{4.10}$$

where κ is a tunable confidence parameter for increasing misclassification confidence toward label l', the actual adversarial example is given by Eq. 4.11.

$$\mathbf{x}' = \frac{1}{2} \left[\tanh(\operatorname{arctanh}(\mathbf{x}) + \delta) + 1 \right] \tag{4.11}$$

where the tanh activation function is used in replacement of box-constraint optimization. For non-targeted attacks, Eq. 4.10 should be updated as:

$$f(\mathbf{x}') = \max \left\{ G(\mathbf{x}')_l - \max_{i \neq l} \{G(\mathbf{x}')_i\}, -\kappa \right\} \tag{4.12}$$

4.2.7 DeepFool Adversarial Attack

Moosavi-Dezfooli *et al.* (Moosavi-Dezfooli, Fawzi & Frossard, 2016) proposed a white-box algorithm for finding the most optimal perturbation for redirecting the position of a legitimate sample toward a pre-defined target label using linear approximation. The optimization problem for achieving optimal δ is given by Eq. 4.13.

$$\arg\min \|\delta\|_2 \quad \text{s.t.} \quad \text{sign}(f(\mathbf{x}')) \neq \text{sign}(f(\mathbf{x})) \tag{4.13}$$

where $\delta = -f(\mathbf{x})\mathbf{w}/\|\mathbf{w}\|_2^2$ and $\mathbf{w}$ is the weight vector. DeepFool can also be modified to a non-targeted attack optimizing for hyperplanes of the victim model. In this paper, we implement targeted DeepFool attack and averaged over available labels measuring over ℓ_2 and ℓ_∞. In practice, this scenario is not only faster but also more destructive than BIMs.

In the next section, we provide a brief overview of common 2D representations of audio signals using time-frequency transformations. Finally, we carry out our adversarial experiments on the transformed audio signals (spectrograms).

4.3 2D Audio Representations

Representing audio signals using time-frequency plots is a standard operation in audio and speech processing representing such signals in a compact and informative way. Fourier transform and wavelet transform are the most commonly used approaches for mapping an audio signal into frequency-magnitude representations. In this section, we briefly review some of the most common approaches.

4.3.1 Short-Time Fourier Transform (STFT)

For a given continuous signal $a(t)$ which is distributed over time, its STFT using a Hann window function $w(\tau)$ can be computed using Eq. 4.14.

$$\text{STFT}\left\{a(t)\right\}(\tau,\omega) = \int_{-\infty}^{\infty} a(t)w(t-\tau)e^{-j\omega t}dt \tag{4.14}$$

where τ and ω are time and frequency axes, respectively. This transform is quite generalizable to discrete-time domain for a discrete signal $a[n]$ as:

$$\text{STFT}\left\{a[n]\right\}[m,\omega] = \sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n} \tag{4.15}$$

where $m \ll n$ and ω is a continuous frequency coefficient. In other words, for generating the STFT of a discrete signal, we need to divide it into overlapping shorter length sub-signals and compute Fourier transform on them, which results in an array of complex coefficients. Calculating the square of the magnitude of this array yields a spectrogram representation as shown in Eq. 4.16.

$$\text{Sp}_{\text{STFT}}\left\{a[n]\right\}[m,\omega] = \left|\sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n}\right|^2 \tag{4.16}$$

This 2D representation shows frequency distribution over discrete-time, and compared to the original signal $a[n]$, it has a lower dimensionality, although it is a lossy operation.

4.3.2 Mel-Frequency Cepstral Coefficients (MFCC)

This transform is a variation of the STFT with some additional postprocessing operations, including non-linear transformation. For every column of the achieved spectrogram, we compute its dot product with several Mel filter banks (power estimates of amplitudes distributed over frequency). For increasing the resolution of the resulting vector, logarithmic filtering should be applied, and finally, it will be mapped to another representation using discrete cosine transform.

This representation has been widely used for sound enhancement and classification. Furthermore, it has been well studied as a standard approach for conventional generative models incorporating Markov chain and Gaussian mixture modes (Shi *et al.*, 2018; Maurya *et al.*, 2018).

4.3.3 Discrete Wavelet Transform (DWT)

Wavelet transform maps the continuous signal $a(t)$ into time and scale (frequency) coefficients similar to STFT using Eq. 4.17.

$$\text{DWT}\left\{a(t)\right\} = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} a(t)\psi\left(\tfrac{t-\tau}{s}\right) dt \tag{4.17}$$

where s and τ denote scale and time variations, respectively, and ψ is the core transformation function known as mother function (see Eq. 4.18). There are a variety of mother functions for different applications, such as the complex Morlet, which is given by Eq. 4.18:

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega t} e^{-t^2/2} \tag{4.18}$$

Discrete-time formulation for this transform is shown in Eq. 4.19.

$$\text{DWT}\left\{a[k,n]\right\} = \sum_{n=-\infty}^{\infty} a[n]\psi\,[n,k] \tag{4.19}$$

where n and k are integer values for the continuous mother function of h. Spectral representation for this transformed signal is a 2D array which is computed by Eq. 4.20:

$$\text{Sp}_{\text{DWT}}\left\{a[n]\right\} = \left|\text{DWT}\left\{a[k,n]\right\}\right|^2 \tag{4.20}$$

In the next section, we explain our experiments on three benchmarking sound datasets. We firstly generate separate spectrogram sets with the three representations mentioned above using different configurations. Second, we train a ResNet on these datasets and run adversarial attack algorithms against them. Finally, we measure both the fooling rate and the cost of attacks.

We demonstrate that for different spectrogram configurations, these metrics are meaningfully different.

4.4 Experiments

We use three environmental sound datasets in all our experiments: UrbanSound8k (Salamon *et al.*, 2014a), ESC-50 (Piczak, 2015b), and ESC-10 (Piczak, 2015b). The first dataset includes 8732 four-second length audio samples distributed in 10 classes: engine idling, car horn, children playing, drilling, air conditioner, jackhammer, dog bark, siren, gunshot, and street music. ESC-50 is a comprehensive dataset with 50 different classes and overall 2000 five-second audio recordings of natural acoustic sounds. A subset of this dataset is ESC-10 which has been released with ten classes and 400 recordings.

For increasing both the quality and the number of samples of these datasets, we apply a pitch-shifting augmentation approach with scales 0.75, 0.9, 1.15, and 1.5 as proposed in (Esmaeilpour *et al.*, 2020a), which positively affect classification accuracy. This data augmentation operation generates four extra audio samples for every original audio sample, and eventually, it increases the size of the original dataset by the factor of four. We discuss the usefulness of this 1D data augmentation approach in Section 4.5. In the following subsection, we explain the details of generating 2D representations for audio signals. Toward this aim, we use the open-source Librosa signal processing python library (McFee, Raffel, Liang, Ellis, McVicar, Battenberg & Nieto, 2015b) and our upgraded version of the wavelet toolbox (Hanov, 2008).

4.4.1 Generating Spectrograms

For every dataset including augmented signals, we separately generate independent sets of 2D representations, namely MFCC, STFT, and DWT. We aim to investigate which audio representation yields a better trade-off between recognition accuracy and robustness for a victim model against various strong adversarial attacks.

4.4.1.1 MFCC Production Settings

There are four major settings in generating MFCC representation using Librosa. The default value for sampling rate is 22.05 kHz. Since there is no optimal approach for determining the best sampling rate, we generate the most informative spectrogram. We run extensive experiments using sampling rates from 8 to 24 kHz. The second tunable hyperparameter is the number of MFCCs (N_{MFCC}), which we examine different values for it: 13, 20, and 40 per frame with a hop length of 1024. Normalization of discrete cosine transform (type 2 or 3) using orthonormal DCT basis for MFCC production is the third setting. By default, this hyperparameter is set to true in almost all the libraries, including Librosa. However, we measure the performance of the front-end classifier trained to MFCC representation without normalization. The last argument is about the number of cepstral filtering (CF) (Juang, Rabiner & Wilpon, 1987) to be applied on MFCC features. The sinusoidal CF reduces involvement of higher-order coefficients and improve recognition performance (Paliwal, 1999) (see Eq. 4.21).

$$\mathbf{M}\left[n,:\right] \leftarrow \mathbf{M}\left[n,:\right] \times \left(1+\sin\left(\frac{\pi(n+1)}{CF}\right)\right)\frac{CF}{2} \tag{4.21}$$

where $\mathbf{M}$ stands for MFCC array with size $[n,:]$. We investigate the effect of CF on the overall performance of classification models.

4.4.1.2 STFT Production Settings

For producing STFT representations, we use default configurations for general hyperparameters as outlined in the Librosa manual. We use 2048, 1024, and 512 with associated sampling rates for assigning the length of the windowed signal. We also use variable window sizes: 2048 (default value), 1024, and 512 (very small window) associated with a default hop size of 512. We investigate the potential effects of these configurations for the resiliency of the victim models against adversarial attacks.

4.4.1.3 DWT Production Settings

For generating DWT representations, we modified the sound explorer software (Hanov, 2008) to support Haar and Mexican Hat wavelet mother functions in addition to complex Morlet. Sampling frequency for DWT spectrograms has been set up to 8 kHz and 16 kHz with a constant frame length of 50 ms. Moreover, by convention, the overlapping threshold is set to 50%. Our experiments measure the impacts of these DWT configurations visualized in logarithmic scale (for higher resolution) on both recognition accuracy and robustness against adversarial attacks.

In the following subsection, we discuss possible choices for the classification models to be separately trained on the spectrogram representations and setups mentioned above. Finally, we select our final front-end classifier from a diverse domain of traditional handcrafted-based feature learning algorithms to state-of-the-art DL architectures.

4.4.2 Classification Model

For the choice of classification algorithms, we initially included both conventional classifiers such as linear and Gaussian SVM (Esmaeilpour *et al.*, 2020), random forest (Esmaeilpour *et al.*, 2020a), and some deep learning architectures. Specifically, we selected pre-trained GoogLeNet (because of its inception mechanism), AlexNet (for taking advantage of its fully convolutional configuration), and ResNet (utilizing a mixture of residual and convolutional blocks) (He *et al.*, 2016) models tuned for our three benchmarking datasets. We preserved the architectures of these ConvNets except for the first layer and the last layer for mapping logits into class labels (softmax layer). Since spectrograms may have different dimensions according to their length and transformation schemes, we bilinearly interpolate them to fit 128×128 for all the ConvNets.

Performance comparison of the SVMs, GoogLeNet, and AlexNet mentioned above against a few adversarial attacks have already been studied mainly for DWT representations of environmental sound datasets in (Esmaeilpour *et al.*, 2020). However, their experiments have been conducted on standard spectrograms without validating the potential impacts of different settings in producing different representations. In this paper, we carry out extensive experiments using: (i) three

common 2D representations for audio signals, namely MFCC (represented in 2D matrix format, not the common vector visualization), STFT, and DWT; (ii) more and stronger targeted and non-targeted algorithms for adversarial attacks; (iii) fair comparison on fooling rates of victim models taking their cost of attacks averaged over the allocated budgets into account.

We primarily select a ConvNet as our front-end classifier for the sake of simplicity and interpretability of results. We present concise results for other classification models in Section 4.5. We selected ResNet architectures for such an aim because such a ConvNet is currently the best-performing classifier for several tasks (Hershey, Chaudhuri, Ellis, Gemmeke, Jansen, Moore, Plakal, Platt, Saurous, Seybold et al., 2017). Our implementations corroborate that on average, these ConvNet architectures outperform all the algorithms mentioned above (both SVMs and other DL approaches) trained on spectrograms. Among the possible architectures for ResNet (ResNet-18, ResNet-34, and ResNet-56), we selected ResNet-18 according to its highest recognition performance and relatively low number of parameters compared to others. Recalling that we investigate the potential effects of spectrogram configurations on the classifier, which has a very competitive recognition accuracy compared to others and requires fewer training parameters. Thus herein, we specifically focus on the ResNet-18 network, and all our investigations will consider this victim architecture.

For every configuration to produce the 2D representations, we generate an individual set of spectrograms and train an independent ResNet-18 classifier on each dataset. We use a 5-fold cross-validation setup on 70% of the overall dataset volume (training plus development). To avoid overtraining, we implemented the early stopping technique in training and finally reported mean recognition accuracy on the test sets (30% remaining).

4.4.3 Adversarial Attacks

In this section, we provide details for attacking the models trained on 2D audio representations. We examine their robustness against six strong adversarial attacks by reporting obtained average fooling rates using the area under the ROC curve (AUC) as a performance metric.

Model robustness refers to the average recognition accuracy of the victim classifier evaluated on the adversarial examples (spectrograms in our case) (Ma *et al.*, 2018). In other words, it measures the ratio of correctly classified adversarial spectrograms over the total number of crafted examples using the AUC metric. It is worth mentioning that there is an inverse relationship between the model robustness and attack fooling rate. More specifically, the latter measures the ratio of misclassified adversarial spectrograms over the total number of crafted examples (see (Ma *et al.*, 2018) for more details).

To the best of our knowledge, all the adversarial attack algorithms are optimization-based procedures toward achieving the minimum possible perturbation. These procedures should generate spectrograms very similar to the ground-truth using a specific similarity metric. This metric is often one of the statistical norms such as l_0, l_2, l_∞, etc. (Goodfellow *et al.*, 2015). Thus, attack algorithms should minimize over the designated similarity metric in an iterative pipeline. The total number of times (in each batch) which this pipeline should be executed until achieving a valid (in terms of being far enough from the decision boundary of the victim model (Papernot, 2018)) adversarial spectrogram is called gradient computation or callback to the ground-truth. This process imposes considerable computational overhead to the entire attack optimization procedure and limits the adversary's strength in runtime. Therefore, increasing the number of required gradient computations is a potential way to decrease the fooling rate of the victim model and potentially resist attacks.

Thus far, it has been demonstrated that the fooling rate of a classifier is dependent on the properties of the attack algorithm, the allocated budget in runtime, and the characteristics of the victim model (Papernot, 2018). For instance, some attack algorithms (e.g., CWA) can get closer to the decision boundary of the victim classifier and consequently find a smaller adversarial perturbation. This results in more effectively attacking the recognition model and increases the fooling rate. Furthermore, since changing the settings of the spectrograms modifies the decision boundary of the audio classifiers, it will most likely affect the fooling rate of the victim model.

4.4.3.1 Settings for Attack Algorithms

In FGSM and BIMs attacks, possible ranges for δ have been defined from 0.001 to any possible supremum under different confidence intervals ($\geq 65\%$). For the implementation of the DeepFool attack, we use the open-source Foolbox package (Rauber *et al.*, 2017) with iterations from 100 to 1000 (10 different scales with a step of 100). In the implementation of the JSMA attack, the number of iterations has been set to $(m_i\gamma)/n_i$ where m_i and n_i denote the total number of pixels and scaling factor within $[0, 200]$ (with displacement a of 40), respectively. Also γ is the maximum allowed distortion (ideally $< 1.5/255$) within the maximum number of iterations. Budget allocated to CWA is within $\{1, 3, 7, 9\}$ for search steps in c within $\{25, 100, 1\text{k}, 2\text{k}, 5\text{k}\}$ iterations in each search step using early stopping. For targeted attacks (i.e., FGSM, JSMA, and CWA) we randomly select targeted wrong labels for running adversarial optimization algorithms.

We executed these attack algorithms on two NVIDIA GTX-1080-Ti with 4×11 GB of memory except for the DeepFool attack, which was executed on 64-bit Intel Core-i7-7700 (3.6 GHz) CPU with 64 GB memory. For attacks on the smallest dataset (ESC-10), we used batches of 200 samples. For larger datasets (ESC-50 and UrbanSound8k), we used 25 batches of 100 samples.

4.4.3.2 Adversarial Attacks for MFCC Representations

We firstly investigate the potential effect of different sampling rates in MFCC production on the performance of the trained models. To this end, sampling rates have been selected from reasonably low (8 kHz) to moderately high (24 kHz) ranges, including the default frequency value (22.05 kHz) defined in Librosa. Therefore, we trained four ResNet-18 models per dataset associated with four sampling rates. The results summarized in Table 4.1 show that the recognition performance of the classifiers is, to some extent, dependent on the sampling rates. For example, for ESC-10 and UrbanSound8k datasets, the sampling rate of 8 kHz improves recognition accuracy, while 16 kHz works better for ESC-50. These results might imply that a low sampling rate filters out high-frequency components and negatively affects the learning of discriminative features from the spectrograms.

Table 4.1 Performance comparison of models trained on MFCC representations with different sampling rates averaged over experiments and budgets. Relatively better performances are in boldface.

Benchmarking Dataset	Sampling Rate (kHz)	Recog. Acc. (%)	AUC Score, Number of Gradients for Adversarial Attacks					
			FGSM	DeepFool	BIM-a	BIM-b	JSMA	CWA
ESC-10	8	**73.23**	**0.9822**, 1	0.9473, 074	**0.9710**, 065	**0.9801**, 110	**0.9308**, 096	**0.9912**, 1346
	16	72.15	0.9456, 1	**0.9607**, 046	0.9334, 059	0.9375, 197	0.9144, 151	0.9616, 1435
	22.05	72.06	0.9467, 1	0.9518, 129	0.9309, 088	0.9379, 186	0.9145, 213	0.9405, 1471
	24	70.13	0.9471, 1	0.9341, 078	0.9298, 115	0.9327, 171	0.9233, 091	0.9302, 1149
ESC-50	8	69.89	0.9517, 1	0.9023, 061	0.9612, 084	0.9703, 193	0.9288, 118	0.9598, 2418
	16	**70.21**	**0.9849**, 1	**0.9912**, 248	**0.9871**, 209	**0.9903**, 160	**0.9508**, 251	0.9672, 2639
	22.05	69.97	0.9534, 1	0.9386, 331	0.9430, 423	0.9581, 288	0.9233, 219	0.9434, 2318
	24	67.25	0.9433, 1	0.9214, 208	0.9307, 159	0.9415, 216	0.9187, 417	**0.9652**, 2744
US8k	8	**71.25**	**0.9905**, 1	**0.9895**, 326	0.9411, 317	**0.9950**, 223	**0.9623**, 398	**0.9708**, 2791
	16	70.81	0.9508, 1	0.9215, 631	0.9346, 519	0.9389, 817	0.9447, 442	0.9449, 3805
	22.05	69.57	0.9457, 1	0.9151, 269	**0.9449**, 184	0.9256, 513	0.9370, 416	0.9456, 3015
	24	69.33	0.9440, 1	0.9221, 318	0.9236, 299	0.9120, 862	0.9242, 343	0.9371, 2816

We attack these models using those six adversarial algorithms mentioned above and measure their fooling rates averaged over different budgets as explained in Section 4.4.3. From the results shown in Table 4.1, we notice an inverse relationship between recognition accuracy and robustness of these models, on average. For instance, ResNet-18 trained on MFCC representation of the ESC-10 dataset sampled at 8 kHz reaches the highest recognition accuracy. Still, this model is less robust against five out of six adversarial attacks, averaged over the allocated budgets. We present two hypotheses on this issue. Firstly, adversarial attacks are essentially optimization-based problems, and their final results are dependent on the hyperparameters defined by the adversary. Confidence intervals, number of callbacks to the original spectrogram, number of iterations in optimization formulation, line search for the optimal coefficient are among those, to name a few. The fooling rate of a victim model is dependent on tuning these hyperparameters. Our second hypothesis is on the statistical perspective of training a neural network. A model with higher recognition accuracy has probably learned a better decision boundary via maximizing the intra-class similarity and inter-class dissimilarity. Attacking this model provides a broader search area for the adversary to find pinholes of the model, especially when the decision boundaries among classes lie in the vicinity of each other. Table 4.1 also compares the average number of gradients for batch execution required by every attack algorithm. Regarding statistics of this table, CWA is the costliest adversarial attack for spectrograms with different sampling rates.

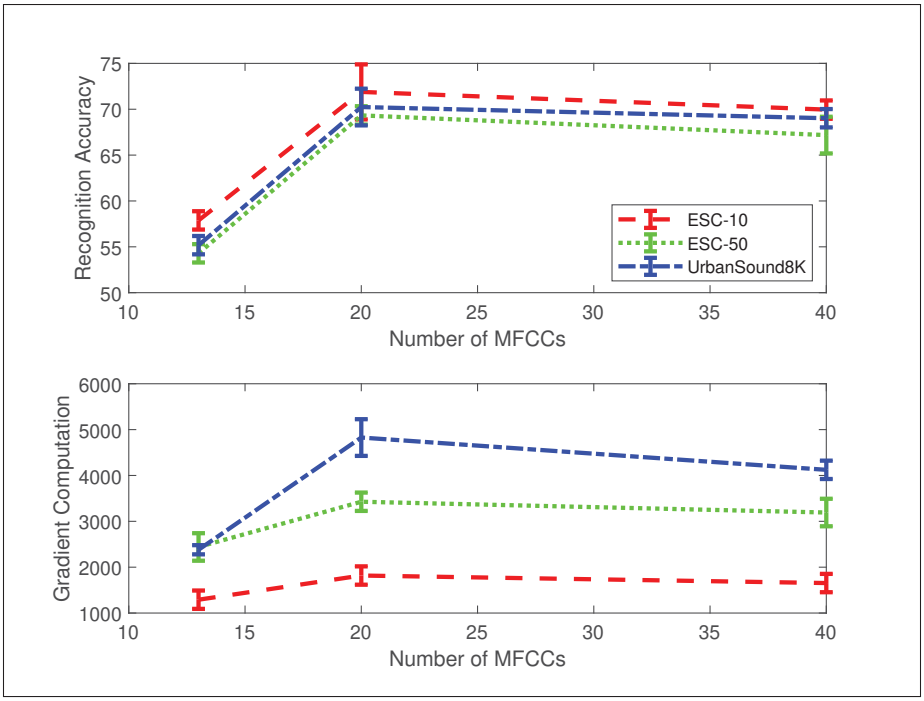

Figure 4.1 Effect of N_{MFCC} on the front-end classifier

The default value for the number of MFCCs (N_{MFCC}) is 20 as defined in Librosa. However, we encompass values from a minimum number of 13 to a maximum of 40 in generating MFCC representation; although increasing N_{MFCC}>20 introduces redundancy in frequency coefficient representation. Our experimental results corroborate the negative effect of a low N_{MFCC} in the performance of the classifiers. More specifically, recognition performance of the trained models on spectrograms with N_{MFCC} = 13 is 14% less than models trained on spectrograms with $N_{\text{MFCC}}\geq$20, on average. Our experimental results on attacking victim models trained on spectrograms with low N_{MFCC} unveils their extreme vulnerabilities. However, in terms of the attack cost, these models need fewer callbacks for gradient computations for yielding AUC>90% (see Figure 4.1). That could be due to the nature of the adversarial attacks, which are formulated as optimization problems, regardless of the performance of the victim models.

Using orthonormal discrete cosine transform basis function is a standard approach in crafting MFCC representation. Our experiments produced two separate subsets of spectrograms with and without normalization to measure its potential effect on recognition accuracy and the fooling rate

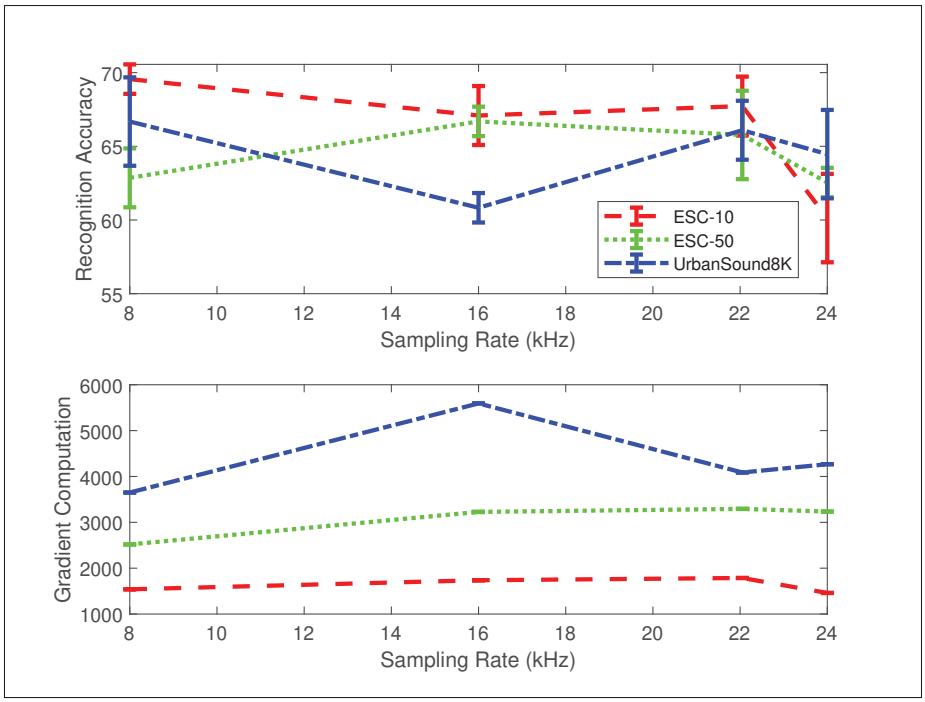

Figure 4.2 Normalization effect on the front-end classifier

(see Figure 4.2). Disabling this normalization scheme causes a drop of 7% in the recognition accuracy and a drop of 8.5% in the attack cost, on average.

For the choice of the cepstral filtering, we covered values in the range $\left[0, (d \times N_{\text{MFCC}})\right]$ where maximum d is 2.5 with hop size of 0.5 in the production of spectrograms. Values above the supremum of this interval generate higher-order coefficients in linear-like weighting distributions which considerably reduce recognition accuracy on average to about 48%. Optimal values for d are 0, 0.5, and 0.3 for ESC-10, ESC-50, and UrbanSound8k, respectively (see Figure 4.3).

4.4.3.3 Adversarial Attacks for STFT Representations

There is a significant similarity in producing MFCC and STFT spectrograms, mainly in terms of transformation and frequency modulation. Therefore, we omit experimental results relevant to measuring the impacts of sampling rates on the robustness of victim classifiers. Nevertheless, fooling rates of ResNet-18 models on STFT representations are similar to MFCC representations.

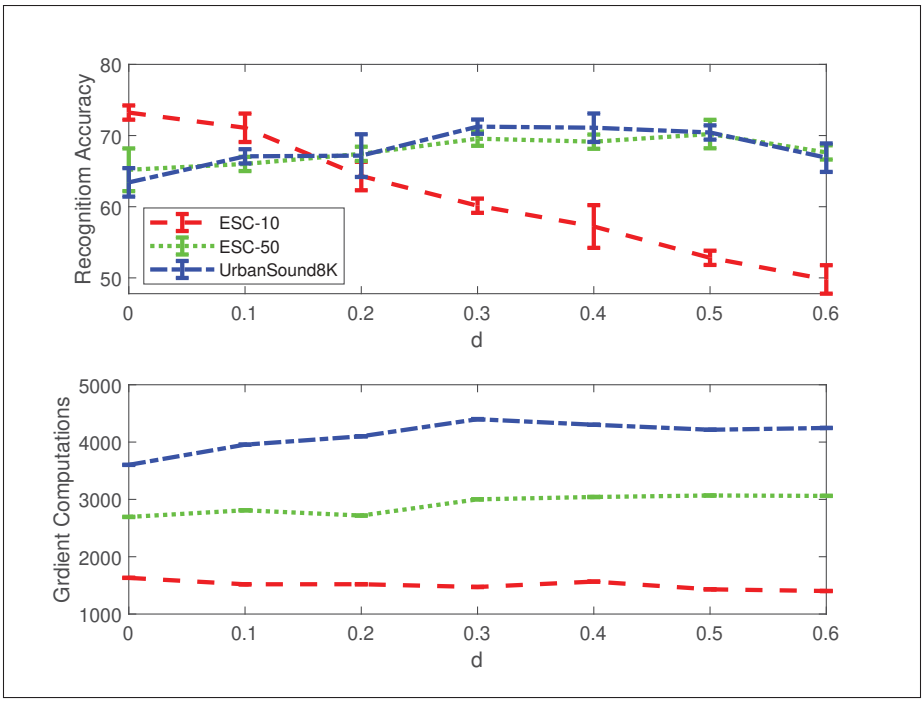

Figure 4.3 Effect of Cepstral filtering on the front-end classifier

Such rates support the inverse relationship between the recognition accuracy and the robustness against attacks mentioned above.

Table 4.2 summarizes adversarial experiments conducted on STFT representations with the same aforementioned setup described in Section 4.4.3. This table illustrates the impact of the number of FFTs (N_{FFT}) both on the recognition accuracy and on the robustness of victim models against adversarial attacks averaged over all the different adversarial setups. For ESC-10 and ESC-50 datasets, N_{FFT}=1024 results in learning better decision boundaries for the classifiers, although it increases fooling rates of the victim models. In the production of STFT spectrograms, each frame of a given audio signal is spanned by a window that covers the frame. The maximum length of this window can be equivalent to the number of N_{FFT}. Since small window lengths improve the temporal resolution of the final STFT representation, we evaluate the performance of the models on small window lengths in the range $\left[\left(0.25\times N_{\text{FFT}}\right), N_{\text{FFT}}\right]$ with hop size of $N_{\text{FFT}}/4$. As shown in Figure 4.4, the evaluation on ESC-50 and UrbanSound8k datasets uncovers that models trained on STFT representations with window length of $0.5\times N_{\text{FFT}}$ outperform others.

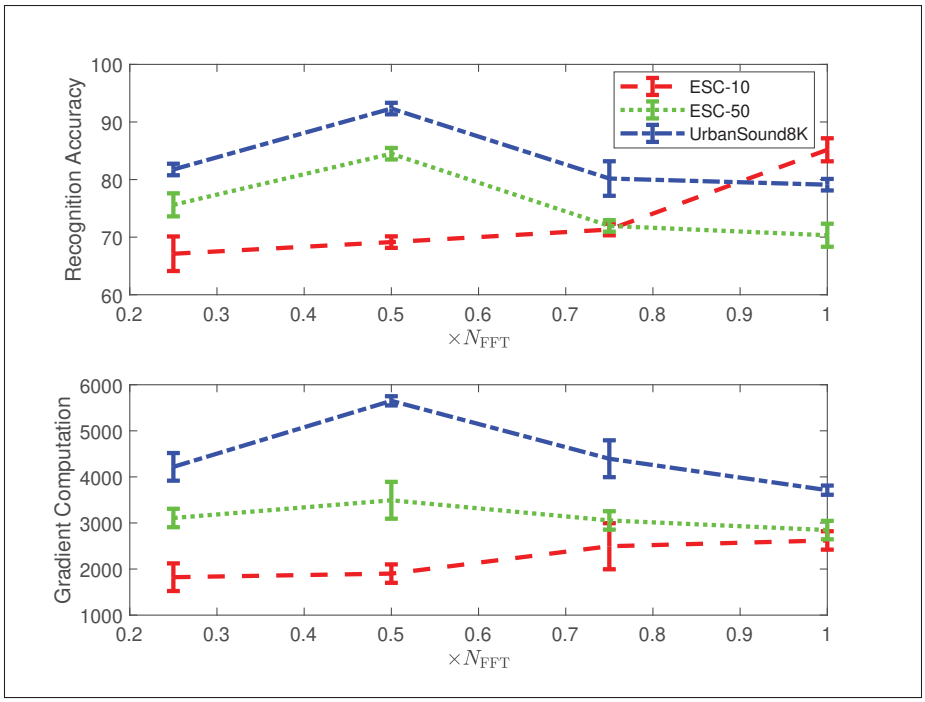

Figure 4.4 Effect of scales for N_{FFT} on the front-end classifier

Table 4.2 Performance comparison of models trained on STFT representations with different N_{FFT} averaged over experiments and budgets. Relatively better performances are in boldface.

Benchmarking Dataset	Number of FFTs	Recog. Acc. (%)	AUC Score, Number of Gradients for Adversarial Attacks					
			FGSM	DeepFool	BIM-a	BIM-b	JSMA	CWA
ESC-10	512	82.41	0.9768, 1	0.9430, 089	0.9576, 109	0.9717, 134	0.9662, 141	0.9846, 1415
	1 024	**85.17**	**0.9823**, 1	**0.9701**, 129	**0.9715**, 091	**0.9792**, 183	0.9531, 209	**0.9905**, 2008
	2 048	80.56	0.9651, 1	0.9544, 092	0.9407, 163	0.9529, 279	0.9588, 341	0.8731, 1730
ESC-50	512	82.44	0.9786, 1	0.9542, 082	0.9583, 109	0.9665, 244	0.9614, 128	0.9618, 1995
	1 024	**84.49**	**0.9881**, 1	0.9512, 331	**0.9871**, 267	**0.9798**, 179	0.9702, 361	**0.9896**, 2353
	2 048	83.12	0.9567, 1	**0.9631**, 145	0.9765, 211	0.9606, 567	**0.9738**, 399	0.9729, 2412
US8k	512	90.58	0.9761, 1	0.9414, 583	**0.9513**, 442	0.9682, 421	0.9402, 345	0.9539, 2569
	1 024	91.74	0.9827, 1	0.9752, 322	0.9340, 471	**0.9687**, 719	0.9515, 502	0.9654, 3271
	2 048	**92.23**	**0.9895**, 1	**0.9764**, 643	0.9407, 602	0.9630, 408	**0.9623**, 655	**0.9673**, 3342

On the ESC-10 dataset, a window length of N_{FFT} resulted in better performance in terms of recognition accuracy.

Comparing the recognition accuracy of Tables 4.1 and 4.2 shows that STFT provides better discriminative features for the ResNet-18 classifier since such a model achieved lower recognition accuracy on MFCC representations. Additionally, while the AUC scores across the six attacks

Table 4.3 Performance comparison of models trained on DWT representations with different sampling rates averaged over different budgets. Relatively better performances are in boldface.

Benchmarking Dataset	Sampling Rate (kHz)	Recog. Acc. (%)	AUC Score, Number of Gradients for Adversarial Attacks					
			FGSM	DeepFool	BIM-a	BIM-b	JSMA	CWA
ESC-10	8	**85.67**	**0.9456**, 1	**0.9310**, 429	0.9307, 612	**0.9411**, 744	**0.9324**, 781	**0.9483**, 4205
	16	82.04	0.9068, 1	0.9192, 672	**0.9437**, 490	0.9347, 513	0.9018, 801	0.9216, 4439
ESC-50	8	80.34	**0.9462**, 1	**0.9335**, 367	0.9161, 452	0.9314, 809	0.9168, 298	0.9233, 3981
	16	**85.97**	0.9376, 1	0.9256, 409	**0.9314**, 628	**0.9419**, 701	**0.9173**, 561	**0.9236**, 4575
US8k	8	**94.70**	**0.9401**, 1	**0.9279**, 761	**0.9315**, 841	**0.9511**, 738	**0.9207**, 691	0.9320, 4684
	16	91.83	0.9321, 1	0.9274, 533	0.9125, 719	0.9408, 941	0.9139, 774	**0.9430**, 4879

are not so different, ranging from 0.93 to 0.99, the number of gradients required for models trained on STFT spectrograms is considerably higher than MFCC representation. In summary, STFT spectrograms provide better accuracy and a little hard to attack, even if they can be fooled with high success by all six adversarial attacks.

4.4.3.4 Adversarial Attacks for DWT representations

There is no algorithmic approach for obtaining the optimal mother function to generate DWT spectrograms. Therefore, from simple Haar to complex Morlet, we have employed several functions to investigate the potential impacts on recognition accuracy and the adversarial robustness of the victim models. We exploited an analytical approach, recasting multiple experiments. Table 4.4 shows that although the complex Morlet mother function outperforms other mother functions in terms of recognition accuracy. However, it shows more vulnerability against adversarial examples, averaged over six attack algorithms with different budgets.

Table 4.3 compares the recognition accuracy of models trained on DWT representations with complex Morlet mother function. We have evaluated these models on DWT spectrograms with sampling rates of 8 kHz and 16 kHz. For ESC-50, a sampling rate of 8 kHz shows better performance for the classifiers, comparing their recognition accuracy. There are three findings in these tables. Firstly, averaged over all the allocated budgets for the attacks, models trained on DWT representations demonstrate slightly higher robustness against adversarial attacks than MFCC and STFT spectrograms. Secondly, the highest recognition accuracy has been achieved

Table 4.4 Comparison of mother functions on the performance of the models. Outperforming values are shown in bold face.

benchmarking Dataset	Mother Function	Average Recognition Accuracy (%)	Average AUC Score
ESC-10	Haar	82.14	95.14
	Mexican Hat	84.51	94.19
	Complex Morlet	**85.67**	**95.61**
ESC-50	Haar	83.08	92.16
	Mexican Hat	84.33	93.40
	Complex Morlet	**85.97**	**95.38**
UrbanSound8k	Haar	91.22	96.16
	Mexican Hat	93.48	95.63
	Complex Morlet	**95.17**	**96.09**

for classifiers trained on DWT representations. Thirdly, the trade-off between recognition accuracy and adversarial robustness of the victim models are noticeable for different sampling rates. Moreover, the cost of the attack (number of gradient computations) for models trained on DWT is considerably higher than the other two representations.

We assumed a frame length of 50 ms with 50% overlapping to convolve the input signal with mother functions in all these experiments. We have also carried out experiments on studying the potential effect of frame length on the performance of the models. They showed that short frame lengths (e.g., 30 ms) drop the recognition performance of the models for the three benchmark datasets. Additionally, long frames such as 50 ms introduce a high redundancy in frequency plots, which results in dropping the recognition accuracy (see Figure 4.5). Figure 4.6 visually compares crafted adversarial examples for the three representations. Although they are visually very similar to their legitimate counterparts, they confidently drive the classifier toward wrong predictions. That showcases the active threat of adversarial attacks for the sound recognition models.

4.5 Discussion

In this section, we provide additional discussion regarding our results. We briefly discuss some secondary aspects of our experiments that could be relevant for future studies.

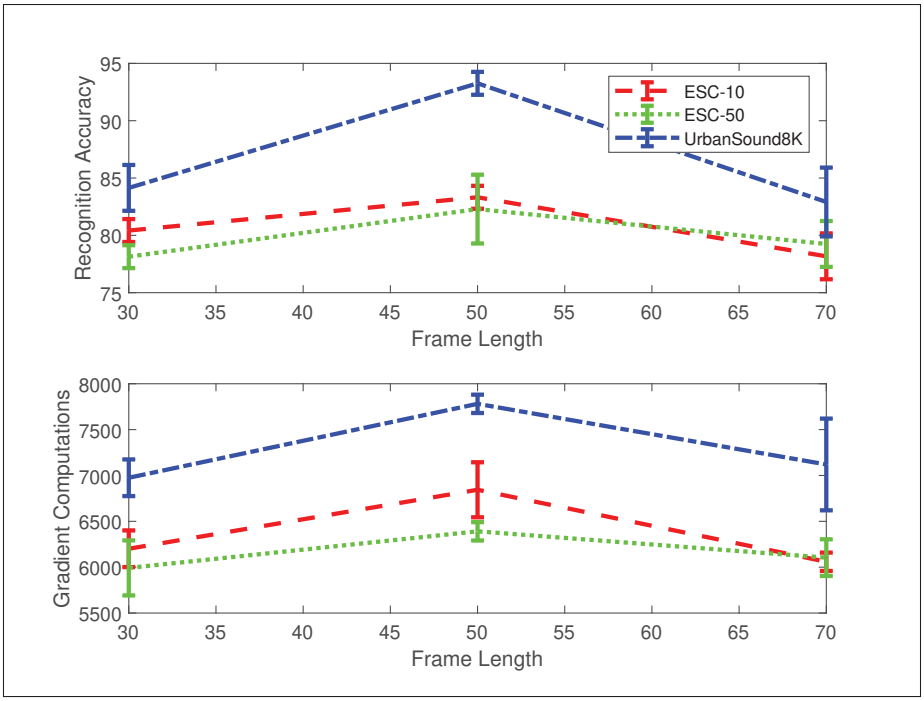

Figure 4.5 The effect of DWT frame length on the front-end classifier

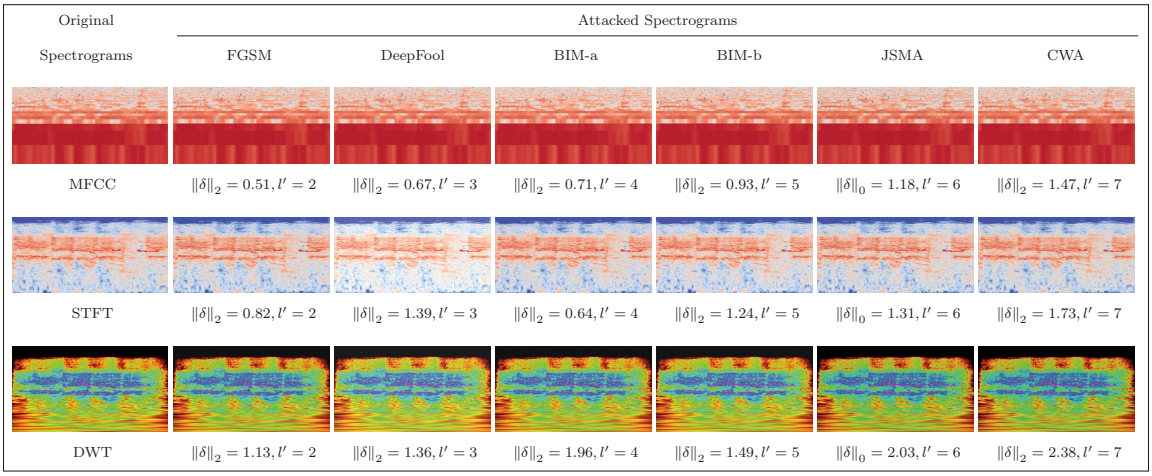

Figure 4.6 Crafted adversarial spectrograms for the three audio representations. The original audio sample has been randomly selected from the class of dog bark ($l = 1$). Examples shown in columns two to seven are associated with the six adversarial attacks for the original input sample. Required perturbation (δ) and the target labels (l') are shown under each spectrogram.

4.5.1 Deep Learning Architectures

We measured recognition accuracy and the total number of training parameters for all candidates for selecting the front-end classifier. We explored DL architectures without residual blocks (AlexNet) and with inception blocks (GoogLeNet) to choose victim classifiers. Our experiments unveiled that these dense networks do not outperform ResNet-18 in terms of recognition accuracy. Although the average recognition accuracy of ResNet-18 and GoogLeNet are competitive on spectrograms, the latter has 1.41× more training parameters. On average, the recognition performance of AlexNet is 8% lower than ResNet-18, even if it has 61% fewer parameters.

Furthermore, the recognition performance of other ResNet models such as ResNet-34 and ResNet-56 are very competitive to ResNet-18, but the latter requires 50% fewer parameters. In comparing the robustness of these models against adversarial attacks, they all can reach fooling rates higher than 95%. Taking the allocated budgets into account, the ResNet-18 is the costliest network in terms of the number of required gradient computations for the adversary, followed by GoogLeNet and AlexNet.

4.5.2 Data Augmentation

For improving the performance of the classifiers, we augmented the original datasets only at waveform level (1D) using time-stretching filter except for DWT representations which we additionally scaled the spectrograms by a logarithmic function. Removing 1D data augmentation negatively affects recognition accuracy of the models with drop ratios of about 0.056%, 0.036%, and 0.029% for MFCC, STFT, and DWT spectrograms, respectively. For measuring the robustness of these models against adversarial examples, we executed attack algorithms on random batches of size 100 among the entire datasets. The experimental results have shown that for reaching the fooling rates as close as the values reported in Tables 4.1 to 4.3, less gradient computation is required mainly for JSMA and CWA attacks.

Figure 4.7 Average transferability ratio of adversarial examples among ConvNets. Higher ratios are shown in boldface.

Dataset	Models	MFCC			STFT			DWT		
		ResNet18	GoogLeNet	AlexNet	ResNet18	GoogLeNet	AlexNet	ResNet18	GoogLeNet	AlexNet
ESC-10	ResNet18	1	**0.672**	0.568	1	**0.713**	0.641	1	0.761	**0.774**
	GoogLeNet	**0.693**	1	0.480	**0.637**	1	0.519	0.646	1	**0.684**
	AlexNet	0.491	**0.521**	1	0.540	**0.562**	1	0.633	**0.701**	1
ESC-50	ResNet18	1	**0.644**	0.519	1	**0.661**	0.609	1	**0.755**	0.732
	GoogLeNet	**0.630**	1	0.531	**0.578**	1	0.569	0.507	1	**0.676**
	AlexNet	0.523	**0.536**	1	0.551	**0.601**	1	0.614	**0.699**	1
US8k	ResNet18	1	0.627	**0.677**	1	0.611	**0.710**	1	**0.714**	0.713
	GoogLeNet	**0.634**	1	0.503	0.563	1	**0.699**	**0.723**	1	0.707
	AlexNet	0.577	**0.583**	1	0.703	**0.735**	1	0.705	**0.678**	1

4.5.3 Adversarial on Raw Audio

Optimizing Eq. 4.1 even for a short 1D audio signal sampled at a low rate is very costly, and they are not transferable while being played over the air (Carlini & Wagner, 2018). Toward addressing this interesting open problem, we trained several end-to-end ConvNets on randomly selected batches of environmental sound datasets. Upon running both targeted and non-targeted attacks against ConvNets, we could reduce the performance of victim classifiers by 30% on average. Interestingly, multiplying the adversarial examples by a small random scalar restored the audio waveforms' correct label. In other words, whereas adversarial spectrograms, 1D adversarial audio waveforms are not resilient against any additional perturbation.

4.5.4 Adversarial Transferability

The transferability of adversarial examples is not only dependent on the classifier but also on audio representations. We investigated this aspect on deep neural networks trained on different spectrograms. Table 4.7 reports the transferability ratios averaged over budgets with batch sizes of 100. Crafted adversarial examples for victim models are less transferable in MFCC representations, while DWT spectrograms have higher transferring rates on average. On the other hand, examples generated in the STFT domain are more transferable compared to MFCC. That may be due to the higher order of information in STFT spectrograms.

Unlike other research works (Esmaeilpour *et al.*, 2020) that have evaluated adversarial transferability among different classifiers to identify the most reliable model considering a black-box attack scenario, we have carried out the transferability experiment to identify the most reliable 2D representation. Therefore, we characterize the impact of 2D representation on the transferability of attacks among different models. In other words, we have demonstrated that classifiers trained on MFCC representations have a lower adversarial transferability ratio than models trained on STFT and DWT.

4.5.5 Selection of Benchmarking Adversarial Attacks

All the attack algorithms evaluated in this paper are comprehensive and still top-notch approaches in generating adversarial examples. They are standard benchmarking approaches in developing defense algorithms since they have a unique objective and technique in finding the most fitting adversarial perturbation. See a relevant discussion in (Jang, Zhao, Hong & Lee, 2019; Hu, Yu, Guo, Chao & Weinberger, 2019b).

4.6 Conclusion

In this paper, we have demonstrated the inverse relationship between recognition accuracy and robustness of ResNet-18 trained on 2D representations of environmental audio signals averaged over the allocated budgets by the adversary. This relation is generalizable to other DL architectures, and this is a common behavior for models trained on spectrograms. Additionally, we showed that our front-end classifier could reach the highest recognition accuracy when trained on DWT representation. Furthermore, attacking this model is, on average, more costly for the adversary compared to models trained on MFCC and STFT representations. That proves the superiority of DWT representation for environmental sound recognition.

Moreover, we have examined the transferability of crafted adversarial examples among AlexNet, GoogLeNet, and ResNet-18 for the three spectrogram representations. According to our results, MFCC representation achieved the lowest transferability ratio, averaged over six different

adversarial attacks. In our future studies, we are decided to investigate this property for networks trained on speech datasets.

CHAPTER 5

MULTI-DISCRIMINATOR SOBOLEV DEFENSE-GAN AGAINST ADVERSARIAL ATTACKS FOR END-TO-END SPEECH SYSTEMS

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper Submitted for Publication to « IEEE Transactions on Information Forensics and Security »
in March 2021.

Abstract

This paper introduces a defense approach against end-to-end adversarial attacks developed for cutting-edge speech-to-text systems. The proposed defense algorithm has four major steps. First, we represent speech signals with 2D spectrograms using the short-time Fourier transform. Second, we iteratively find a safe vector using a spectrogram subspace projection operation. This operation minimizes the chordal distance adjustment between spectrograms with an additional regularization term. Third, we synthesize a spectrogram with such a safe vector using a novel GAN architecture trained with Sobolev integral probability metric. To improve the model's performance in terms of stability and the total number of learned modes, we impose an additional constraint on the generator network. Finally, we reconstruct the signal from the synthesized spectrogram and the Griffin-Lim phase approximation technique. We evaluate the proposed defense approach against six strong white and black-box adversarial attacks benchmarked on DeepSpeech, Kaldi, and Lingvo models. Our experimental results show that our algorithm outperforms other state-of-the-art defense algorithms both in terms of accuracy and signal quality.[44]

[44] The supplementary materials including speech signals are available at:
https://github.com/EsmaeilpourMohammad/MultiDiscriminator-SDGAN.git.

5.1 Introduction

There is a large volume of publications on applying deep learning algorithms for audio and speech classification (i.e., transcription), which report high recognition accuracy (Wang, Zou, Chong & Wang, 2020; Shen *et al.*, 2019; Mozilla-DeepSpeech, 2017). During the last decade, the primary focus has been designing new architectures, for instance, variants of convolution (Sainath *et al.*, 2013), recurrent (Graves *et al.*, 2013), and attention configurations (Bahdanau *et al.*, 2016) to improve classification accuracy and model generalizability. However, it has been proven that these advanced models might undergo extreme vulnerability against carefully crafted adversarial signals both in 1D and 2D representation (spectrogram - mainly for audio) domains (Carlini & Wagner, 2018; Esmaeilpour *et al.*, 2020).

The major focus of this paper is in response to this vulnerability issue. We have developed an adversarial defense approach against varieties of end-to-end speech-to-text attack algorithms. Toward this end, we firstly review the state-of-the-art of adversarial attacks in Section 5.2. We also provide details about the background of the defense approaches in Section 5.3. Section 5.4 introduces the proposed adversarial defense algorithm followed by comprehensive experimental results in Section 5.5. In summary, we make the following contributions in this paper:

1. introducing a novel adversarial defense approach based on a multi-discriminator generative adversarial network (GAN) in the restricted Sobolev space (Brezis, 2010);
2. establishing simple yet effective architectures for both the generator and discriminator networks;
3. developing an adjusted chordal distance with a complementary regularization term toward achieving a safe input vector for the generator model;
4. characterizing a constraining technique for improving the stability of our generative model in adverse environmental scenarios;
5. experimentally proving the effectiveness of the proposed defense approach for white and black-box as well as targeted and non-targeted attack scenarios.

5.2 Background: Adversarial Attack

An adversarial signal $\vec{x}_{adv}$ carries inaudible perturbation δ, and it can fool the victim classifier (the transcription model) toward any target phrase $\hat{\mathbf{y}}$ defined by the adversary (Carlini & Wagner, 2018). The actual value of δ is dependent on the length of $\hat{\mathbf{y}}$ (the number of characters - tokens) and the characteristics of the original carrier signal $\vec{x}_{org}$ ($\vec{x}_{adv} = \vec{x}_{org} + \delta$) (Carlini & Wagner, 2018; Qin *et al.*, 2019). For measuring the loudness (distortion) of this perturbation relative to the carrier signal, a logarithmic-scale metric has been proposed by Carlini and Wagner (Carlini & Wagner, 2018):

$$l_{\text{dB}}(\vec{x}_{adv}) = l_{\text{dB}}(\delta) - l_{\text{dB}}(\vec{x}_{org}) \tag{5.1}$$

where $l(\cdot)$ denotes the loudness of the original 1D signal $\vec{x}_{org} \in \mathbb{R}^{n \times m}$ in dB, and n and m denote the length and number of channels, respectively. For $l_{\text{dB}}(\vec{x}_{adv}) < \epsilon$ where ϵ is a small threshold, $\vec{x}_{adv}$ sounds almost seamless to $\vec{x}_{org}$ according to the C&W attack for the speech-to-text model (Carlini & Wagner, 2018):

$$\min |\delta|_2^2 + \sum_i c_i . \mathcal{L}_i(\vec{x}_{org,i} + \delta_i, \pi_i) \quad \text{s.t.} \quad l_{\text{dB}}(\vec{x}_{adv}) < \epsilon \tag{5.2}$$

where c_i is a scaling coefficient for the connectionist temporal classification loss function $\mathcal{L}(\cdot)$ (Graves, Fernández, Gomez & Schmidhuber, 2006). Additionally, π_i denotes string tokens without duplication, which should reduce to the character alignments $\hat{\mathbf{y}}_i$ ($\hat{\mathbf{y}}_i \neq \mathbf{y}_i$, where the latter refers to the ground truth character alignment) (Carlini & Wagner, 2018). The C&W attack has been primarily developed for the speech-to-text DeepSpeech model (Mozilla-DeepSpeech, 2017), and the experiments have shown a complete collapse of this model against adversarial signals crafted through Eq. 5.2 (Carlini & Wagner, 2018).

The C&W attack splits the input signal into 50 frames per second, and it eventually yields a universal perturbation for the entire chunks in $\vec{x}_{org}$. This operation reduces the computational complexity of the attack algorithm compared to optimizing fine-grained δ_i for every chunk. However, it might negatively affect the robustness of $\vec{x}_{adv}$ in a real-world environment. In other words, playing these speech chunks over the air and recording them by another microphone,

involving environmental reverberating and signal echo, might easily remove the adversarial effect (δ) (Schönherr, Eisenhofer, Zeiler, Holz & Kolossa, 2020). Several algorithms crafting more resilient adversarial signals in natural environments have been proposed in response to this issue. These algorithms are based on psychoacoustic loss function (Szurley & Kolter, 2019), feature vector analysis (Abdullah, Garcia, Peeters, Traynor, Butler & Wilson, 2019), and employing a set of filters (band-pass, impulse response, and white Gaussian noise) (Yakura & Sakuma, 2018). However, these approaches have been evaluated within static environments with predefined room setups, which might reduce these algorithms' generalizability in more challenging scenarios (Schönherr *et al.*, 2020). Inspired by Athalye, Engstrom, Ilyas & Kwok (2018a), which introduces the expectation over transformation (EOT) in the attack optimization formulation for regularizing the cost function (similar to Eq. 5.2), many other EOT variants have been proposed for the speech domain (Qin *et al.*, 2019; Schönherr *et al.*, 2020; Chen, Shangguan, Li & Jamieson, 2020). These regularizations help to craft more robust adversarial signals for non-static environments, which fit in both white and black-box attack scenarios.

The EOT proposed by Qin *et al.* (Qin *et al.*, 2019) is based on an acoustic room simulator, which generates artificial utterances and environmental reverberations. This algorithm is known as Robust Attack and encodes the EOT regularization into the loss function of a speech-to-text model as (Qin *et al.*, 2019):

$$\ell(\vec{x}_{org,i}, \delta_i, \mathbf{y}_i) = \mathbb{E}_{t\sim\tau}\left[\ell_{net}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) + \alpha\ell_m(\vec{x}_{org,i}, \delta_i)\right] \tag{5.3}$$

where α is a static scaling factor, $\ell_{net}(\cdot)$ denotes the cross entropy loss and $\ell_m(\cdot)$ indicates the loss function for masking threshold (ϵ). In fact, $\ell_m(\cdot)$ constrains over the normalized power spectral density function of $\vec{x}_{org}$ and contributes to the imperceptibility of the adversarial signal (Qin *et al.*, 2019). Additionally, τ refers to the transformation set including room reverberation settings. This attack has been tested on the Lingvo speech-to-text system (Shen *et al.*, 2019) and it has achieved a very high fooling rate on this advanced system.

The Imperio attack proposes another variant of EOT, which implements simulated room impulse response (RIR) filters, taking advantage of a simple deep neural network (DNN) architecture (Schönherr *et al.*, 2020). Additionally, this attack embeds psychoacoustic thresholding for reducing adversarial distortion similar to Qin *et al.* (Qin *et al.*, 2019) (see Eq. 5.4 in (Schönherr *et al.*, 2020)).

$$\vec{x}_{adv} = \arg\max_{\vec{x}_i} \mathbb{E}_{h \sim H_{\text{dim}}} \left[P(\hat{\mathbf{y}}_i | \vec{x}_{i,h}) \right] \tag{5.4}$$

where $h \in H_{\text{dim}}$ denotes a RIR filter and dim indicates the dimension of the filter set. The Imperio is an iterative algorithm and minimizes the adversarial perturbation δ via approximating the $\nabla_{\vec{x}_{org}} = \partial \ell_{net}(\mathbf{y}, \hat{\mathbf{y}}) / \partial f^*(\vec{x}_{org})$ where $f^*(\cdot)$ denotes the post activation function. In each iteration and according to the distribution of H_{dim}, an adversarial candidate $\vec{x}_{adv} = \vec{x}_i + \kappa \nabla_{\vec{x}_i}$ with the learning rate κ should satisfy $\hat{\mathbf{y}}_i \neq \mathbf{y}_i$. This procedure continues until a predefined audible threshold ϵ has been reached. This attack was evaluated on the Kaldi speech-to-text system (Povey *et al.*, 2011), which employs both DNN and hidden Markov model (HMM) configurations for real-time speech transcription. It has been shown that under various environmental settings, including lecture, meeting, and office rooms, the Imperio attack has considerably turned down the transcription performance of the Kaldi system (Schönherr *et al.*, 2020).

The EOT regularization in the Metamorph adversarial attack (Chen *et al.*, 2020) is similar to the RIR filtration of the Imperio algorithm with one major difference: it implements channel impulse response (CIR) to characterize potential over the air distortions on δ. This attack algorithm employs M pairs of microphone-speaker transmission in different distances (similar to H_{dim}) to encompass a wide range of reverberations in yielding minimal perturbation:

$$\arg\min_{\delta} \alpha_t l_{\text{dB}}(\vec{x}_{adv}) + \frac{1}{M} \mathcal{L}(\vec{x}_{org} + \delta_i, \pi_i) \tag{5.5}$$

where α_t denotes a trade-off scalar between the fooling rate of the model and the signal quality. Similar to the C&W attack, the Metamorph attack was evaluated on the DeepSpeech model. The experiments showed an attack success rate of around 90% and low Mel-cepstral distortion for this white-box algorithm (Chen *et al.*, 2020).

Since integrating the EOT regularization into the adversarial optimization problem requires access to the victim model's cost function, it cannot be directly incorporated in the black-box attacks. For addressing this issue, a surrogate technique has been proposed, called the over-the-line approach. This technique provides multiple varieties of the adversarial signals to the victim model before playbacks over the air (Abdullah *et al.*, 2019). This operation helps the adversary to capture the environmental scene distribution without directly simulating it through reverberation filters. However, the performance of this approach is directly dependent on the comprehensiveness of the over-the-line adversarial signals. More straightforward yet effective black-box adversarial attacks, which do not incorporate EOT regularization with to some extent better performance on the DeepSpeech system, are the genetic algorithm attack (GAA) (Taori, Kamsetty, Chu & Vemuri, 2019) and multi-objective optimization attack (MOOA) (Khare, Aralikatte & Mani, 2019). These algorithms were tested for targeted and non-targeted attacks and achieved high fooling rates.

While all the aforementioned adversarial attacks pose major security concerns against cutting-edge speech-to-text models, namely DeepSpeech, Kaldi, and Lingvo, there are few investigations on defense algorithms. The following section reviews the state-of-the-art defense approaches developed for counteracting white and black-box adversarial attacks.

5.3 Background: Adversarial Defense

Developing defense approaches against robust adversarial attack algorithms can be very challenging due to several reasons. Firstly, standard speech signals have high dimensionality (e.g., 8 kHz), and even running effective compression techniques (Das *et al.*, 2018) for potentially discarding adversarial perturbations can be time-consuming in real-time speech-to-text transcription. Secondly, speech signals often have various channels for quality enhancement purposes (Peinado & Segura, 2006). Thus an adversary can optimize δ for such channel(s), which human auditory systems are less sensitive to them and more effectively fool the victim model (Virag, 1999). Thirdly, speech signals usually carry environmental and microphone-speaker noises, which makes distinguishing a noisy signal from an adversarial one very difficult, even

after band-pass filtering (Hu & Loizou, 2007). In the following, we briefly review a couple of multiscale approaches that have been able to tackle these challenges to some extent.

Inspired by Das, Shanbhogue, Chen, Hohman, Chen, Kounavis & Chau (2017), a compression-based approach has been introduced for removing the potential adversarial perturbation from the speech signals (Das *et al.*, 2018). This algorithm implements both adaptive multi-rate and MPEG audio layer-3 encoding for such an aim. Reported results showed the effectiveness of this approach in adverse scenarios for short-length signals (Das *et al.*, 2018). Furthermore, for sophisticated adversarial signals, which have been precisely optimized through running the Robust Attack (Qin *et al.*, 2019), this defense scheme failed to remove adversarial perturbations (Esmaeilpour, Cardinal & Koerich, 2021a).

An autoencoder-based defense GAN (A-GAN) (Latif *et al.*, 2018) is structurally similar to the compression approach mentioned above. Instead of low-level signal filtering, it implements high-level feature transformation. The intuition behind this approach is transforming the signal into a similar recording with lower environmental noises using an autoencoder. The proposed autoencoder implements a complex architecture for reconstructing feature vectors so as to remove potential adversarial perturbation δ. Extensive experiments of A-GAN on DeepSpeech and Lingvo systems have been reported by Esmaeilpour *et al.* (Esmaeilpour *et al.*, 2021a).

Since it has been proven that adversarial subspace is distinct from original and noisy signals (Esmaeilpour *et al.*, 2020b), a defense GAN based on this fact has been developed by Esmaeilpour *et al.* (Esmaeilpour *et al.*, 2021a). Unlike the compression approach and A-GAN approaches, this defense algorithm employs neither low nor high-level transformations for discarding adversarial perturbations directly in the signal. Instead, it uses a class-conditional GAN for computing a refined latent variable $\mathbf{z}_i$ for the generator network via:

$$\nabla_{\mathbf{z}_i} \|\gamma\left[G(\mathbf{z}_i), \mathbf{x}_i\right]\|_2^2 \tag{5.6}$$

where $\mathbf{z}_i \in \mathbb{R}^{d_z}$ with dimension d_z is the random variable from $p_z \sim \mathcal{N}(0, 0.4I)$ and $G(\cdot)$ with distribution p_g denotes the generator network. Additionally, $\gamma[\cdot]$ is the chordal distant

Algorithm 5.1 $\gamma[\cdot]$ computation. We refer to Appendix I for more details (taken from Esmaeilpour *et al.* (2020b)).

Algorithm: Computing vector of $\gamma[\cdot]$
Input: C_{leg}: Class of legitimate samples
Output: $\gamma[\cdot]$

$\Lambda_{leg} = [] \, , \Lambda_{adv} = []$; /* legitimate and adversarial lists */
for *all* $B_{leg} \in C_{leg}$; /* B_{leg}: legitimate batch */
do
  B_{adv} := adversarial attack on B_{leg}; /* B_{adv}: adversarial batch */
  $\overrightarrow{\lambda}_{leg} = \text{eigen}\left[\text{qz}\left(B_{leg}\left[i\right], B_{leg}\left[j\right]\right)\right]$; /* qz decomposition, $i \neq j$ */
  $\overrightarrow{\lambda}_{adv} = \text{eigen}\left[\text{qz}\left(B_{adv}\left[i\right], B_{adv}\left[j\right]\right)\right]$; /* $i \neq j$ */
  Λ_{leg}.append $\left(\overrightarrow{\lambda}_{leg}\right)$, Λ_{adv}.append $\left(\overrightarrow{\lambda}_{adv}\right)$
end for
$\gamma[\cdot] = \text{DIF}(\Lambda_{leg}, \Lambda_{adv})$; /* $\gamma[\cdot]$ denotes difference. */

adjustment function between the input spectrogram $\mathbf{x}_i$ and $G(\mathbf{z}_i)$ (see Algorithm 5.1 which is driven from Algorithm I-1). Eq. 5.6 is iterative and finds the optimal latent variable $\mathbf{z}_i^*$, which not only forces $G(\mathbf{z}_i^*)$ to lie in the original signal subspace, but also generates a spectrogram very similar to $\mathbf{x}_i$.

The effectiveness of this class-conditional defense GAN (CC-DGAN) has been evaluated against the C&W attack, the Robust Attack, and the GAA for both DeepSpeech and Lingvo systems (Esmaeilpour *et al.*, 2021a). However, it might fail for long-length signals (above six seconds) due to the generator network's instability in around 10k iterations. For addressing this issue, we propose two techniques: (i) introducing a multi-discriminator GAN to provide more informative gradients to the generator network; (ii) implementing such a GAN in the restricted Sobolev space (Brezis, 2010) and training the generator network according to the Sobolev function class with a bounded dominant measure. Since a special case of this restricted space is proportional to the 2D Fourier transform representation (spectrogram) (Brezis, 2010), we can train our generative model in a much lower dimensionality compared to 1D speech signals. In the following section, we explain these steps as part of the proposed defense scheme.

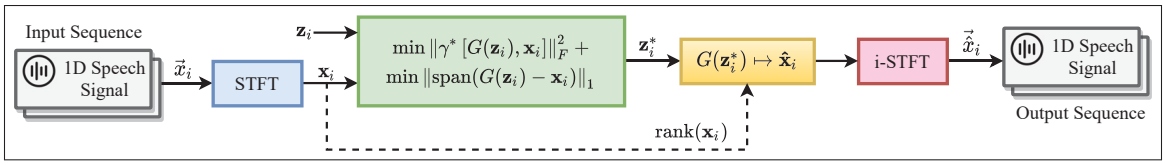

Figure 5.1 An overview of the proposed defense GAN approach. The 1D speech signal ($\vec{x}_i$) is converted to a STFT spectrogram ($\mathbf{x}_i$). Moreover, $\gamma\,[\cdot]$ denotes the chordal distance adjustment required for making $\mathbf{x}_i$ in the same subspace of the synthesized spectrogram $G(\mathbf{z}_i)$ ($\mathbf{z}_i \in \mathbb{R}^{d_z}$ is the latent random variable). The output speech signal ($\vec{\hat{x}}_i$) is reconstructed using the i-STFT operation and the Griffin-Lim phase approximation approach (Masuyama *et al.*, 2019). Additionally, $\text{rank}(\mathbf{x}_i)$ refers to the input spectrogram's rank according to its eigenvalues computed in the Schur decomposition domain.

5.4 Proposed Adversarial Defense Method: Sobolev Defense GAN (Sobolev-DGAN)

The proposed adversarial defense approach against speech attacks has four steps, as depicted in Fig. 5.1: (a) signal representation (conversion from 1D vector to 2D matrix) using short-time Fourier transform (STFT) (Griffin & Lim, 1984); (b) chordal distance adjustment with a complementary regularization term for projecting the given input spectrogram onto the original subspace (the process shown in the green color); (c) spectrogram synthesis using a Sobolev GAN and an optimal safe vector $\mathbf{z}_i^*$ (yellow block in Fig. 5.1); (d) inverse STFT (i-STFT) for reconstructing the speech signal.

5.4.1 Spectrogram: 2D Representation of 1D Speech Signal

There are several standard transformations in the audio and speech processing domains for representing a signal into a 2D spectrogram, such as continuous or discrete wavelet transform, Mel-frequency cepstral coefficients, and STFT. All these transformations have some advantages over each other, and they have been widely used for unsupervised, weakly supervised, and supervised learning tasks. Moreover, the highest recognition accuracies have been often reported for the models trained on these representations over 1D signals (Mozilla-DeepSpeech, 2017; Chorowski, Weiss, Bengio & van den Oord, 2019). This is presumably due to the lower dimensionality of spectrograms and the inherent ability of these transformations in extracting more distinctive learning features compared to 1D signals (Deng & O'Shaughnessy, 2018).

This paper uses the STFT to generate spectrograms from the given speech signals since it is more closely related to the Sobolev integral probability metric (IPM) (Mroueh, Li, Sercu, Raj & Cheng, 2018), which we employ to train our generator network. This metric correlates well with the Fourier coefficients encoded in the STFT spectrograms and likely helps extract more distinctive features. The theoretical approach for crafting an STFT spectrogram is as follows.

For a given discrete signal $a[n]$ with length n (sampled from a 1D speech signal $\vec{x}$ in the time domain), we can define the Fourier transform using a Hann function $A[\cdot]$ as (Griffin & Lim, 1984):

$$\text{STFT}\left\{a[n]\right\}[k,\omega] = \sum_{n=-\infty}^{\infty} a[n]A[n-k]e^{-j\omega n} \tag{5.7}$$

where k is the shifting scale ($k \ll n$) and ω indicates the frequency coefficients. For capturing more features from $a[n]$, this operation applies on the overlapping signal chunks (i.e., 50 ms) according to a predefined sampling rate (e.g., 16 kHz). For generating the spectrogram, we need to compute the power spectrum of Eq. 5.7 as:

$$\text{SP}_{\text{STFT}}\left\{a[n]\right\}[k,\omega] = \left|\sum_{n=-\infty}^{\infty} a[n]A[n-k]e^{-j\omega n}\right|^2 \tag{5.8}$$

where it generates a 2D matrix for a given speech signal $\vec{x}_i$. In the next subsection, we explain the second step of the proposed defense approach, which finds a refined $\mathbf{z}_i^*$ from the combination of a random $\mathbf{z}_i \in \mathbb{R}^{d_z}$ and the original input spectrogram ($\mathbf{x}_i$).

5.4.2 Chordal Distance Adjustment for Spectrogram Projection

In a big picture, there are two categories in developing defense approaches against adversarial attacks:

1. running low or high-level transformations for filtering the input signal aiming at discarding potential adversarial perturbation (as discussed in Section 5.3)

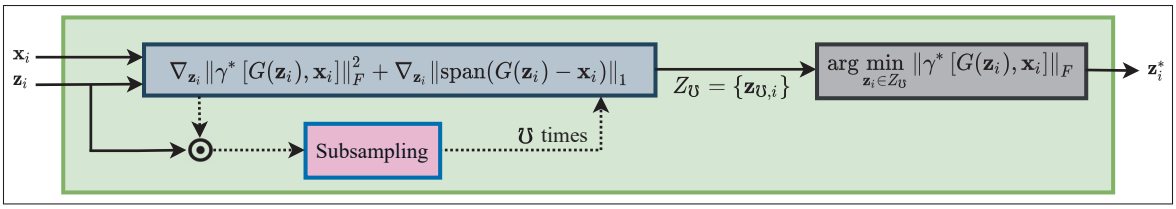

Figure 5.2 Overview of the proposed spectrogram subspace projection using the chordal distance adjustment and a complementary regularization term. The subsampling process is implemented with the distribution $\mathcal{N}(0.5, 0.5I)$ (ratio of 0.5) for avoiding ill-conditioned pencils (Van Loan & Golub, 1983), and a dotted line shows the internal loop. Upon producing a candidate set of $Z_{\mho}$ vectors from the given inputs, we select that $\mathbf{z}_i$ which minimizes the adjusted chordal distance between the synthesized spectrogram $G(\mathbf{z}_i)$ and the input spectrogram $\mathbf{x}_i$.

2. synthesizing a very similar signal to a given input vector without running any filtration operation (Esmaeilpour *et al.*, 2021a; Samangouei, Kabkab & Chellappa, 2018a).

While most of the introduced algorithms fall into the first category, they are often less reliable since they obfuscate gradient vectors (Athalye *et al.*, 2018b). However, developing a synthesis-based defense algorithm is more challenging since it requires two key steps — a projection of the input space and a stable generative model. Since the proposed defense approach fits the second category, therefore we introduce novel techniques for these steps.

The main goal in this step is finding a safe $\mathbf{z}_i^* \in \mathbb{R}^{d_z}$ for the generator network according to two main conditions: $G(\mathbf{z}_i^*)$ should lie in the subspace of the original signal distribution represented by p_r (approximated by p_g); the synthesized spectrogram $G(\mathbf{z}_i^*)$ should be very similar to the spectrogram of the given 1D speech signal ($\mathbf{x}_i$) using the ℓ_2 distance metric. Toward this end, for every input spectrogram $\mathbf{x}_i$, we solve an optimization problem searching all possible $\mathbf{z}_i \in \mathbb{R}^{d_z}$ to find the $\mathbf{z}_i^*$ that meets the conditions above. Fig. 5.2 shows an overview of this operation.

Inspired by Xingjun *et al.* (Ma *et al.*, 2018), which proved that adversarial examples lie in distinct subspaces from original and noisy input samples, the chordal distance metric has been introduced for measuring interspaces among spectrogram manifolds (Esmaeilpour *et al.*, 2020b). This metric, defined in the Schur decomposition domain for the triplet of original, noisy, and

adversarial spectrograms, can be written as (Van Loan & Golub, 1983):

$$\text{chord}(\lambda[G(\mathbf{z}_i)], \lambda[\mathbf{x}_i]) \leq \frac{\epsilon}{\sqrt{\left[\left(\Phi^{\mathcal{H}} G(\mathbf{z}_i)\Gamma\right) + \left(\Phi^{\mathcal{H}} \mathbf{x}_i \Gamma\right)\right]^2}} \tag{5.9}$$

where $\epsilon \leq 20\text{dB}$ is the maximum audible perturbation threshold, which can be defined (or optimized) by the adversary, $\lambda[\cdot]$ denotes the eigenvalue vector function class obtained with Schur decomposition. Γ, Φ, and $\Phi^{\mathcal{H}}$ (conjugate transpose of Φ) are random unit 2-norm operators, which satisfy (Van Loan & Golub, 1983):

$$\mathbf{x}_i\Gamma = \lambda[\mathbf{x}_i] G(\mathbf{z}_i)\Gamma \quad \text{and} \quad \Phi^{\mathcal{H}}\mathbf{x}_i = \lambda[G(\mathbf{z}_i)]\Phi^{\mathcal{H}} G(\mathbf{z}_i) \tag{5.10}$$

For simplicity, we assume that these operators are static for all samples. Although this assumption simplifies the computation, it might result in ill-conditioned cases where an adjustment $\gamma[\cdot]$ is needed ($\text{chord}(\cdot) + \gamma[\cdot]$) (Van Loan & Golub, 1983). It has been shown that this adjustment is relatively large for adversarial spectrograms compared to original and noisy samples (Esmaeilpour *et al.*, 2020b). Therefore, iteratively minimizing over $\gamma[\cdot]$ for $\text{chord}(\lambda[G(\mathbf{z}_i)], \lambda[\mathbf{x}_i])$ considerably increases the chance of finding the safe $\mathbf{z}_i^*$ that satisfies the conditions mentioned above (Esmaeilpour *et al.*, 2021a, 2020b).

Since $\lambda[\cdot]$, defined in the Schur decomposition domain, is sorted (descending) and it is inductive (coefficient of both $\lambda[G(\mathbf{z}_i)]$ and $\lambda[\mathbf{x}_i]$ have upper bound (Brezis, 2010)), according to the Zorn lemma (Brezis, 2010) there exists a relative maximal coefficient for both $G(\mathbf{z}_i)$ and $\mathbf{x}_i$ in the Hahn–Banach analytic form. Thus, we define:

$$\underbrace{\gamma[\lambda[G(\mathbf{z}_j)], \lambda[\mathbf{x}_j]]}_{\gamma^*[\cdot]} \leq \gamma[\lambda[G(\mathbf{z}_i)], \lambda[\mathbf{x}_i]] \quad \text{for} \quad j \ll i \tag{5.11}$$

where j should be chosen according to the properties of the spectrograms. However, we empirically set $j \doteq \max(i) \cdot 0.25$ to make a reasonable trade-off between spectrogram quality and computational complexity (75% improvement). On the other hand, this operation might

constitute ill-conditioned pencils (a pencil is a manifold in the closed-form of $\psi G(\mathbf{z}_i) - \mathbf{x}_i$ where $\psi \propto p_g$ (Van Loan & Golub, 1983)) by discarding $(i - j)$ eigenvectors. To tackle this challenge, we add a complementary regularization term to the spectrogram subspace projection formulation:

$$\nabla_{\mathbf{z}_i} \|\gamma^* [G(\mathbf{z}_i), \mathbf{x}_i]\|_F^2 + \underbrace{\nabla_{\mathbf{z}_i} \|\text{span}(G(\mathbf{z}_i) - \mathbf{x}_i)\|_1}_{\text{regularization}} \tag{5.12}$$

where span$(\cdot)$ computes a linearly independent manifold in the Schur decomposition domain from the difference between the input and synthesized spectrograms (Van Loan & Golub, 1983). The intuition behind this regularization term is tying $G(\mathbf{z}_i)$ as close as possible to $\mathbf{x}_i$ and counteracting with the potential ill-conditioned pencils imposed from $\gamma^*[\cdot]$. Ill-conditioned cases often happen when $\gamma^*[\cdot]$ is minimized, but $G(\mathbf{z}_i)$ and $\mathbf{x}_i$ are not similar.

Upon solving this optimization problem (Eq. 5.12), we achieve a candidate set $Z_{\mho} = \left\{\mathbf{z}_{\mho,i}\right\}$ among all the possible $\mathbf{z}_i \in \mathbb{R}^{d_z}$. Finally, we find the most optimal vector from $Z_{\mho}$ via solving for:

$$\mathbf{z}_i^* := \arg \min_{\mathbf{z}_i \in Z_{\mho}} \|\gamma^* [G(\mathbf{z}_i), \mathbf{x}_i]\|_F \tag{5.13}$$

where $\mathbf{z}_i^*$ is presumably refined to provide a safe input vector for the generator model. We do not directly filter the spectrograms to remove adversarial perturbation δ. We find a reliable vector for a generative model to synthesize a similar spectrogram. However, the performance of all these operations is highly dependent on the generalizability and stability of the GAN model.

5.4.3 Spectrogram Synthesis Using a Sobolev-GAN

The generative model proposed for synthesizing spectrograms is based on the vanilla GAN (Goodfellow *et al.*, 2014) but with an integral probability metric defined in the Sobolev space (Brezis, 2010; Mroueh *et al.*, 2018). Since a specific case of such a space correlates with the Fourier transform, we use this measure for training our GAN on STFT spectrograms. Moreover, we introduce novel architectures for both generator and discriminator networks. For improving

the generalizability and the stability of the entire model, we propose imposing a constraint on the restricted Sobolev space and incorporating multiple discriminator networks.

The task of a generator network in a GAN configuration is minimizing the discrepancies between the synthesized (p_g) and real/original (p_r) sample distributions based on a specific measure (Goodfellow *et al.*, 2014). The choice of such a measure is quite important since it contributes to the generalizability of the entire model (both generator and discriminator networks) (Arjovsky & Bottou, 2017). During the last years, many improvements have been made in designing comprehensive distance measures on top of the φ-divergence (Goodfellow *et al.*, 2014) such as Wasserstein (Arjovsky & Bottou, 2017), Stein (Feng, Wang & Liu, 2017), Cramér (Bellemare *et al.*, 2017), maximum mean discrepancy (MMD) (Dziugaite, Roy & Ghahramani, 2015; Li, Chang, Cheng, Yang & Póczos, 2017), and μ-Fisher IPM (Mroueh & Sercu, 2017). The function which measures this discrepancy is called critic, and it can be formulated (in the closed-form) as (Müller, 1997):

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}_{G(\mathbf{x}_i) \sim p_g} f(G(\mathbf{x}_i)) - \mathbb{E}_{\mathbf{x}_{org} \sim p_r} f(\mathbf{x}_{org}) \right] \tag{5.14}$$

where $\mathcal{F}$ refers to the function class, which is independent of p_g and p_r (Sriperumbudur, Fukumizu, Gretton, Schölkopf, Lanckriet et al., 2012). For improving the GAN stability during training, restriction often applies to the critic function following the characteristics of $\mathcal{F}$ such as Lipschitz continuity ($\|f\|_{\text{Lip}} \leq 1$) in Wasserstein-GAN (Arjovsky & Bottou, 2017) and kernel Hilbert unit ball ($\|f\|_{\text{Hil}} \leq 1$) in MMD-GAN (Li *et al.*, 2017). Moreover, these restrictions should be in line with the properties of the training sample modality. They might result in a weak or unstable generative model, especially for sequence generation (e.g., text and speech) (Mroueh *et al.*, 2018).

The similarity measure used for training our GAN is the Sobolev IPM, adapted for sequence-to-sequence generation (Mroueh *et al.*, 2018) such as chunks of speech signals. Formally, the function class in the Sobolev space with the zero boundary condition and the dominant

probability density function $\mu(\cdot)$ has the following definition (Brezis, 2010; Mroueh *et al.*, 2018):

$$\mathcal{F} = \left\{ \mathcal{X} \to \mathbb{R}^{d_z} \mid\mid \mathcal{X} \to L^{p_s}(\mathbb{R}^{d_z}), f \in W^{k_s,p_s}(\mathcal{X}, \mu), \mathbb{E}_{\mathbf{x}\sim\mu} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \leq 1 \right\} \tag{5.15}$$

where $\mu \sim \mathbb{P}(p_r, p_g), k_s, p_s \in \mathbb{N}$. Additionally, $\mathcal{X} \in \mathbb{R}^{d_z}$ is a compact open subset, $L^{(\cdot)}$ indicates the Lebesgue norm for $1 \leq p_s \leq \infty$, k_s denotes the order of the critic function, and $\mathbb{P}$ is the probability density function. The special case of the function class $\mathcal{F}$ is for $p_s = 2$ where it forms a Hilbert space $\mathcal{H}^{k_s} = W^{k_s,2}$ in connection with Fourier transform as follows (Brezis, 2010):

$$\mathcal{H}^{k_s}(\cdot) \simeq \sum \alpha_s \left| \hat{f}(x) \right|^2 < \infty, f \in L^2(\cdot) \tag{5.16}$$

where α_s is a scalar, and $\hat{f}(x)$ refers to the Fourier series for $f(\cdot)$. Since a spectrogram is also a set of Fourier coefficients, $W^{k_s,2}$ provides a meaningful domain for capturing local distributions of $\mathrm{SP}_{\mathrm{STFT}}$. We also assume $k_s = 1$ and simplify the underlying Sobolev space as (Mroueh *et al.*, 2018):

$$W^{1,2}(\mathcal{X}, \mu) = \left\{ f : \mathcal{X} \to \mathbb{R}^{d_z}, \int_{\mathcal{X}} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \mu(\mathbf{x}) d\mathbf{x} < \infty \right\} \tag{5.17}$$

where this restricted Sobolev space also constraints the critic function f into a unit ball $\mathbb{E}_{\mathbf{x}\sim\mu} \left\| \nabla_{\mathbf{x}} f(\mathbf{x}) \right\|^2 \leq 1$. There are numerous possible choices for defining the dominant measure $\mu(\cdot)$ according to this restricted space's properties. However, we initialize it to $\mu(\cdot) = 0.5 \cdot (p_r + p_g)$ which is the optimal case in training a GAN (Mroueh *et al.*, 2018). Based on these explanations and using Eq. 5.14, we can formulate the Sobolev GAN as (Mroueh *et al.*, 2018):

$$\min_{G_{\theta_g}} \left[\sup_{f_\vartheta, \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla_{\mathbf{x}} f_\vartheta(\tilde{\mathbf{x}}) \right\|^2 \leq 1} \right] \mathcal{E}(f_\vartheta, G_{\theta_g}) = \frac{1}{N} \left(\sum_{i=1}^{N} f_\vartheta(G(\mathbf{z}_i)) - \sum_{i=1}^{N} f_\vartheta(\mathbf{x}_i) \right), \vartheta \geq 2 \tag{5.18}$$

where the critic function f_ϑ follows the imposed constraint in Eq. 5.17, and ϑ is the degree of the critic function. Additionally, N refers to the total number of training samples, and θ_g denotes the weight vectors of the generator network. Moreover, for supporting the continuity and smoothness of f_ϑ, especially for higher-order ϑ, it is recommended to define (Gulrajani

et al., 2017):

$$\tilde{\mathbf{x}}_i = \alpha_\vartheta \left(u\mathbf{x}_{r,i} + (1-u)G(\mathbf{z}_i) \right) \tag{5.19}$$

where $u \sim \mathcal{U}[0,1]$ and α_ϑ is an empirical hyperparameter (we initialize it to $\alpha_\vartheta = 0.9$). This change of variable implicitly interpolates between p_r and p_g to enhance generator model stability (Gulrajani *et al.*, 2017). However, this enhancement is also dependent on the configurations of both the generator network in optimizing for Eq. 5.18 and the discriminator network, which provides gradient vectors to G_{θ_g}.

Our proposed architecture for the generator network employs convolution and residual blocks due to their representation power in capturing continuous density functions of the input space (Radford, Metz & Chintala, 2016; Brock *et al.*, 2019) such as spectrograms (see Fig. 5.3). The generator network contains a fully connected 1D vector layer equivalent to the total dimension of the spectrogram (128×128), followed by batch normalization (BN) and the rectified linear unit activation function (ReLU). This network's first hidden layers are two convolution blocks with the receptive field and stride of 5×5×1. The second hidden layer contains three consecutive residual blocks where each of them has a dilated convolution operation with aggregation. Inspired by Kumar *et al.* (Kumar, Kumar, de Boissiere, Gestin, Teoh, Sotelo, de Brébisson, Bengio & Courville, 2019), the filter sizes of these blocks are identical. Finally, this network's output layer is a transposed convolution (Mao *et al.*, 2018), which yields an RGB spectrogram.

Since the discriminator network provides gradients to the generator and has a crucial role in the entire model's stability (Miyato, Kataoka, Koyama & Yoshida, 2018), we empirically embedded five discriminators with identical architectures. However, we unloaded these networks from residual and long short-term memory (LSTM) blocks to avoid unnecessary complications. The filter sizes in these networks are different, and they escalate by a factor of two so as to encompass a broader range of spectrum distribution. Unlike the generator network, all the convolution layers in the discriminators deploy leaky ReLU (LReLU), as discussed in (Zhang *et al.*, 2017).

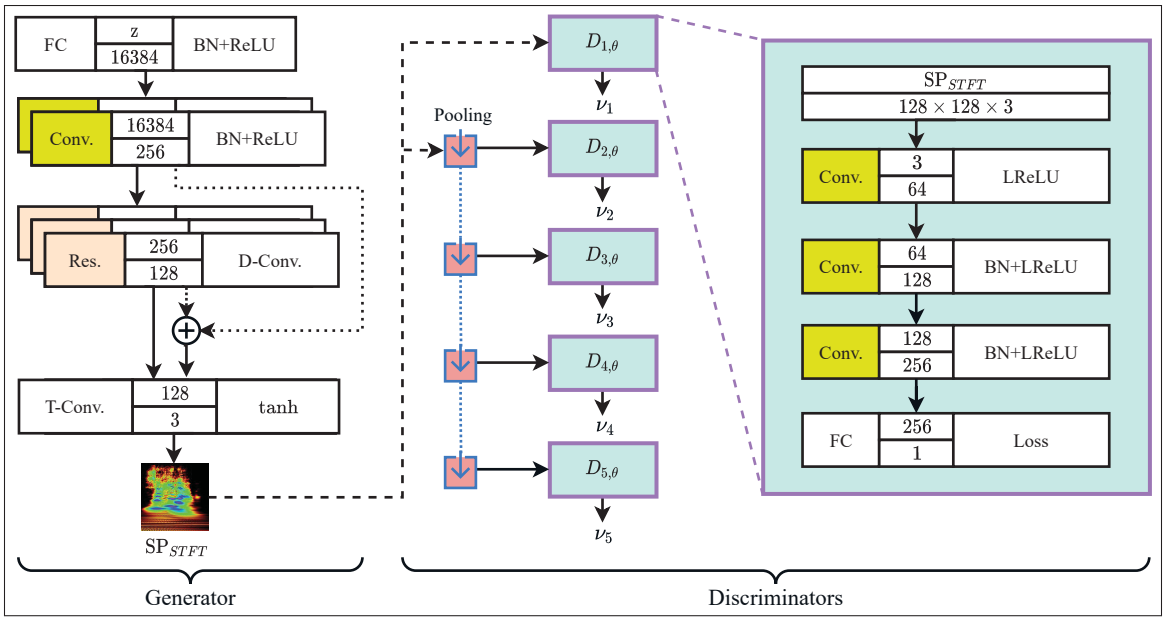

Figure 5.3 Overview of the proposed GAN architecture (one generator and five discriminators $D_{i,\theta}$ for $\forall i = 1:5$) for spectrogram synthesis. Fully connected (FC), convolution (Conv.), dilated convolution (D-Conv.), transposed convolution (T-Conv.), and residual (Res.) layers are followed by weight normalization. The top and bottom parts of the layers refer to the input and output filters' dimensions, respectively. Moreover, ν_i for $\forall i = 1:5$ denotes the logits of the discriminator.

The general formulation for training these GANs is:

$$\min_{G} \max_{D_i} \mathbb{E}_{\mathbf{x}\sim p_r} \left[\log D_i(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_z} \left[\log\left(1 - D_i(G(\mathbf{z}))\right)\right] \tag{5.20}$$

where $\forall i = 1:5$ and $p_{\mathbf{z}} \sim \mathcal{N}(0, I)$. The loss function of these networks is similar to the hinge objective function introduced in (Miyato *et al.*, 2018). However, according to the Sobolev IPM:

$$\mathcal{L}_S(\vartheta, \theta_g, \theta_d, \varrho_s) = \mathcal{E}(f_\vartheta, G) + \varrho_s(1 - \Omega_s(f_\vartheta, G)) - \frac{\rho}{2}(\Omega_s(f_\vartheta, G) - 1)^2 \tag{5.21}$$

and in this definition, $\Omega_s(\cdot)$ in the restricted Sobolev space $W^{1,2}(\mathcal{X}, \mu)$ is differentiable and regarding Eq. 5.18, it is defined as (Mroueh *et al.*, 2018):

$$\frac{1}{2N}\left(\sum_{i=1}^{N} \|\nabla_{\mathbf{x}} f_\vartheta(\mathbf{x}_i)\|^2 + \sum_{i=1}^{N} \|\nabla_{\mathbf{x}} f_\vartheta(G(\mathbf{z}_i))\|^2\right) \tag{5.22}$$

Moreover, ϱ_s, θ_d, and $\rho > 0$ denote the Lagrange multiplier, the weight vectors of each discriminator network, and the penalty weight for providing higher smoothness in training, respectively (Mroueh & Sercu, 2017). One potential side effect of training the generator with multiple discriminators is the difficulty of making a trade-off between sample variety and quality. For tackling this challenge, we use orthogonal regularization (OR) for all the discriminator networks using a simple linear similarity measure (Brock, Lim, Ritchie & Weston, 2017):

$$R_{\varpi} = \varpi \left\| \theta_d^{\top} \theta_d \odot (\mathbf{1} - I) \right\|_F^2 \tag{5.23}$$

where empirically $\varpi \in (10^{-5}, 10^{-4}]$ is a small tuning coefficient, and $\mathbf{1}$ indicates a matrix with constant values of one (Brock *et al.*, 2019). This regularization forces the discriminator network to reduce dissimilarity among filters to learn more distinctive features. However, this might negatively affect the generator performance in capturing all the possible modes from the spectrogram, cause instability in a higher number of iterations, and generate oversmoothed samples (Esmaeilpour, Cardinal & Koerich, 2020c). In response to this issue, we propose a new constraint for the critic function f_{ϑ} as the following.

Proposition: There is an achievable upperbound (supremum) for the continuous (and partially differentiable) critic function $f_{\vartheta}(\cdot)$ in the restricted Sobolev space $W^{1,2}(\mathcal{X}, \mu)$ with:

$$L^{\eta}(\mathcal{X}) = \left\{ f_{\vartheta} : \mathcal{X} \to \mathbb{R}, |f_{\vartheta}|^{\eta} \in L^1(\mathcal{X}) \right\} \tag{5.24}$$

where $\|f\|_{L^1} = \|f\|_1$ and $1 \leq \eta \leq \infty$. This reduces the space definition in Eq. 5.17 to $\int_{\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \mu(\mathbf{x}) d\mathbf{x} \leq c_{\Upsilon}$ where c_{Υ} is a positive static scalar.

Proof: According to the rigid constraint $\mathbb{E}_{\mathbf{x}\sim\mu} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \leq 1$ imposed on $W^{1,2}(\mathcal{X}, \mu)$ in Eq. 5.17, it always supports $\|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \in L^{\eta}$ (the Lebesgue norm). If we bind $\mu(\mathbf{x}) \in L^{\eta'}$ where η' denotes the conjugate exponent of η ($1/\eta + 1/\eta' = 1$), then using the Hölder's inequality (Brezis,

2010), we can write:

$$\int_{\mathcal{X}} \|\nabla_{\mathbf{x}} f(\mathbf{x})\|^2 \, \mu(\mathbf{x}) d\mathbf{x} \leq \underbrace{\|\nabla_{\mathbf{x}} f(\mathbf{x})\|_{\eta}^2 \, \|\mu(\mathbf{x})\|_{\eta'}}_{c_{\Upsilon} << \infty} \quad \square \tag{5.25}$$

where c_{Υ} is dependent on the cumulative distribution of $\mu(\mathbf{x})$. This constraint forces the generator network to discard local sample distributions which lie far from the optimal generator distribution ($\lceil p_r + p_g \rceil / 2$). It also implicitly helps the discriminator network avoid shattering gradients vectors since the learning space is bound to c_{Υ}.

For synthesizing a spectrogram similar to the given $\mathbf{x}_i$, the generator network maps the safe vector $\mathbf{z}_i^*$ onto $\hat{\mathbf{x}}_i$ and then tunes the generated spectrogram with the $\mathbf{x}_i$'s rank (Van Loan & Golub, 1983) in the Schur decomposition domain. Even if this tuning is optional, it improves the quality of $\hat{\mathbf{x}}_i$ and reduces the potential dissimilarity between $G(\mathbf{z}_i^*)$ and $\mathbf{x}_i$.

The last step of the proposed adversarial defense approach is transforming the synthesized spectrograms into the time domain using the inverse STFT operation. This step is necessary only for end-to-end speech-to-text victim models upon adversary's discern.

Reconstructing an audio or speech signal from a spectrogram requires the associated phase vectors from the transformation function (e.g., STFT). There are two main approaches for such an aim: using original phase vectors and approximating phase vectors. Obviously, in the first approach, the reconstructed signals' quality will be very similar to the original counterparts since they share the same timing. However, original phase vectors might not always be accessible, contrary to the second signal reconstruction approach. On the other hand, approximated phase vectors usually add audible noise to the reconstructed signal and degrade its quality. Therefore we opted for the second approach since accessing the original phase vectors might be prohibitive in some senses. Specifically, we use the recognized Griffin-Lim algorithm for the i-STFT procedure (Masuyama *et al.*, 2019). Since this may raise concerns about the quality of the reconstructed signals, we measure their peculiarity with some metrics.

5.5 Experimental Results

In this section, we analyze the proposed defense scheme's performance from two points of view: the defense algorithm's success rate by measuring the word error rate and sentence-level accuracy scores, and the quality of the signals from the synthesized spectrograms and the approximated phase vectors. The latter also includes comparing signals after filtration by various defense algorithms. This shows the impact of defense algorithms on speech signals.

Our benchmarking victim models are DeepSpeech, Kaldi, and Lingvo, which employ both the conventional and cutting-edge learning blocks, such as HMM, convolutional, recurrent, LSTM, and residual configurations. These models are trained on Mozilla common voice (MCV) (MCV, 2019) and LibriSpeech (Panayotov *et al.*, 2015) comprehensive datasets, including numerous utterances. Moreover, they contain above 1,000 hours of recordings organized in short ($\leq$ 6 sec) and long ($>$ 6 sec) voice clips.

In all our experiments, we use a combination of strong white and black-box end-to-end adversarial attacks, as discussed in Section 5.2. For every adversarial signal, regardless of EOT type, we assign ten targeted incorrect different phrases, including silence (Carlini & Wagner, 2018), and five non-targeted incorrect random phrases with different lengths to more effectively challenge defense approaches. Meanwhile, we take identical assumptions for those algorithms that require environmental settings such as CIR and RIR filter sets for fairness in comparison. Following a common practice in adversarial studies (Carlini & Wagner, 2018; Qin *et al.*, 2019; Taori *et al.*, 2019), we also craft adversarial signals for a group of randomly selected portions (with shuffling) of the datasets mentioned above. More specifically, we randomly choose 25k English-speaking samples from both MCV and LibriSpeech with an almost equal number of genders (male and female), accent (e.g., United States, England, etc.), and age (the majority between 19 to 39 regarding the dataset limitation). We assign almost 60% of these samples for training, tuning, and validating our generative model. Hence, the remaining portion will be used for developing adversarial signals using six attack algorithms.

Since we train our GAN on the spectrograms, we firstly convert speech signals into SP_{STFT} with a sampling rate of 22.05 kHz. Additionally, we set the total number of Mel-frequency coefficients to 20 per frame with an overlapping ratio of 0.5 and the hop length of 512. The Hann window length is initialized to 2048 with reflect padding.

We discard checkpoints with unstable learning curves during training and opt to early stop when any signs of instability become present (Brock *et al.*, 2019). For all the architectures (the generator and five discriminators) we use the Adam optimizer with a static learning rate of $2 \cdot 10^{-5}$ and hyperparameters $\beta_1 = 0$ and $\beta_2 = 0.9$. We empirically set the required number of steps for the generator network over the discriminators to two with a decay ratio of 0.99 on four NVIDIA GTX-1080-Ti and two 64-bit Intel Core-i7-7700 (3.6 GHz) with 8×11GB and 2×64GB memory, respectively.

For evaluating the performance of the proposed defense algorithm against adversarial attacks, we also use the word error rate (WER) and sentence level accuracy (SLA) (Qin *et al.*, 2019). The first metric measures the summation of total phrase insertion, substitution, and deletion over the ground-truth phrases ($\mathbf{y}_i$). The second metric measures the ratio of correctly transcripted phrases over the total number of test speech signals. To avoid bias in our analysis, we repeat each experiment 10 times and report the average WER and SLA for each defense algorithm. Table 5.1 summarizes the achieved results.

Table 5.1 shows that for most cases, the proposed defense approach (Sobolev-DGAN*) and its variant without employing the constraining proposition (Sobolev-DGAN) introduced in Section 5.4.3 outperform other defense algorithms against six strong end-to-end speech attacks. Averaged over all the conducted experiments on the three victim speech-to-text models, Sobolev-DGANs have similar performance on white (C&W, Metamorph, Imperio, and Robust Attack) and black-box (GAA and MOOA) attack algorithms. That indicates the independence of our defense algorithm to the adversarial attack scenarios. Moreover, the total number of required iterations ($\mho$) toward achieving the safe input vector $\mathbf{z}_i^*$ for the C&W attack and the Robust Attacks is relatively more than others. That could be interpreted as the higher power of these

Table 5.1 Comparison of the defense algorithms against strong white and black-box adversarial attacks for the DeepSpeech, Kaldi, and Lingvo victim speech-to-text models. Unlike WER and LLR, higher values for the SLA, PESQ, segSNR, and STOI metrics are better. The difference between Sobolev-DGAN* and Sobolev-DGAN is the latter does not incorporate the constraint proposition (Eq. 5.25) mentioned in Section 5.4.3. Outperforming results are shown in boldface.

Model	Attack	Defense	Average ʊ	WER (%)	SLA (%)	PESQ	segSNR	STOI	LLR
DeepSpeech	C&W	Compression	–	19.14 ± 2.36	49.26 ± 2.67	1.64	09.31	0.85	0.44
		A-GAN	–	26.32 ± 3.03	36.21 ± 0.12	1.15	06.95	0.87	0.41
		CC-DGAN	–	14.52 ± 1.16	61.23 ± 1.02	2.01	12.56	0.89	0.38
		Sobolev-DGAN	163	07.61 ± 0.47	76.15 ± 2.18	2.36	18.73	0.91	0.31
		Sobolev-DGAN*	159	**04.21 ± 1.39**	**79.24 ± 1.17**	**2.71**	**19.91**	**0.95**	**0.30**
	Metamorph	Compression	–	21.54 ± 2.17	51.57 ± 1.91	1.55	10.34	0.76	0.48
		A-GAN	–	19.81 ± 3.72	58.39 ± 0.49	1.59	10.86	0.83	0.32
		CC-DGAN	–	11.89 ± 1.23	71.94 ± 1.56	1.96	11.08	0.85	0.35
		Sobolev-DGAN	039	09.37 ± 1.12	75.19 ± 2.18	2.17	14.76	0.88	0.34
		Sobolev-DGAN*	027	**06.79 ± 0.19**	**80.34 ± 3.67**	**2.45**	**16.01**	**0.93**	**0.31**
	GAA	Compression	–	27.41 ± 3.61	43.71 ± 1.32	2.14	14.37	0.87	0.39
		A-GAN	–	29.49 ± 5.26	40.88 ± 5.37	1.66	12.53	0.88	0.37
		CC-DGAN	–	14.98 ± 3.56	69.46 ± 2.37	2.03	13.52	0.90	0.34
		Sobolev-DGAN	101	09.68 ± 2.73	**73.98 ± 0.77**	**2.39**	16.02	0.93	**0.29**
		Sobolev-DGAN*	097	**05.01 ± 0.11**	72.88 ± 4.28	2.38	**18.91**	**0.94**	0.30
	MOOA	Compression	–	17.06 ± 0.19	55.16 ± 3.86	1.87	**19.42**	**0.92**	**0.38**
		A-GAN	–	18.74 ± 43.21	53.07 ± 3.06	1.85	14.63	0.87	0.41
		CC-DGAN	–	15.69 ± 1.97	61.11 ± 2.99	1.99	17.81	0.89	0.39
		Sobolev-DGAN	051	12.25 ± 2.84	68.84 ± 1.56	**2.46**	19.35	0.90	0.36
		Sobolev-DGAN*	049	**04.23 ± 2.32**	**79.36 ± 2.16**	2.30	18.06	0.91	0.39
Kaldi	Imperio	Compression	–	16.29 ± 5.17	56.42 ± 6.11	**2.42**	15.79	0.83	**0.32**
		A-GAN	–	17.76 ± 0.16	54.28 ± 1.90	1.23	09.76	0.74	0.48
		CC-DGAN	–	10.19 ± 2.93	69.62 ± 2.63	1.84	**16.53**	0.78	0.45
		Sobolev-DGAN	093	06.78 ± 0.91	75.33 ± 2.97	1.96	13.98	0.81	0.41
		Sobolev-DGAN*	047	**03.29 ± 1.14**	**82.37 ± 3.62**	2.35	16.52	**0.89**	0.35
Lingvo	Robust Attack	Compression	–	21.56 ± 4.15	55.11 ± 3.05	2.06	15.08	0.74	0.33
		A-GAN	–	17.90 ± 4.21	59.98 ± 1.38	**2.17**	14.43	0.72	0.34
		CC-DGAN	–	14.46 ± 0.35	64.16 ± 2.14	1.71	11.09	0.79	0.28
		Sobolev-DGAN	114	11.99 ± 2.76	69.33 ± 0.81	1.92	12.25	0.76	0.34
		Sobolev-DGAN*	136	**05.86 ± 1.64**	**83.46 ± 2.27**	1.96	**17.07**	**0.81**	**0.22**

attacks in yielding more destructive adversarial signals since they demand an additional cost for our defense algorithm to find the input vector. However, any discussion on the resiliency of adversarial attacks and their potentials in optimizing upscale examples is beyond this paper's scope.

Furthermore, Table 5.1 also proves the effectiveness of the proposed constraining technique for the critic function f_ϑ as discussed in Section 5.4.3. Except for the GAA, Sobolev-DGAN* has shown higher SLA than the Sobolev-DGAN on all the victim speech-to-text models.

For evaluating the potential negative impact of running defense algorithms on the crafted adversarial signals, we use four objective speech quality metrics: perceptual evaluation of

speech quality (PESQ) (Rix, Beerends, Hollier & Hekstra, 2001), segmental signal to noise ratio (segSNR) (Baby & Verhulst, 2019), short-term objective intelligibility (STOI) (Taal, Hendriks, Heusdens & Jensen, 2011), and log-likelihood ratio (LLR) (Baby & Verhulst, 2019). The first metric is based on cognitive modeling, and the input filter set aligns with identifying noisy intervals (high-level quality analysis). The second metric is the enhanced version of the conventional signal-to-noise ratio in audible logarithmic scale for chunks of speech signals (low-level quality analysis). The third metric evaluates the ratio of band-pass local noise perceptibility to the entire signal chunks. Unfortunately, these metrics are not normalized in a scaled interval. However, there is a direct relationship between their magnitudes and signal quality. The fourth metric is associated with a logarithmic noise ratio relative to the ground-truth scaled between $[0, 1]$. Therefore, high-quality signals have lower LLR. As shown in Table 5.1, for the most cases, averaged over ten times experiment repetitions, both the Sobolev-DGAN* and Sobolev-DGAN outperform others in keeping the quality of the signals after running the defense filtration.

In Section 5.4.3, we mentioned that $W^{k_s,2}$ provides a meaningful (and comprehensive) domain for capturing local distributions of spectrograms. To investigate this claim, Fig. 5.4 shows the relation between the Sobolev IPM and extracted local and global probability distributions from spectrograms compared to others. Toward this end, inspired by Mao *et al.* (Mao *et al.*, 2018), we compare the mode collapse issue between the GANs trained with various IPMs as mentioned in Section 5.4.3. We have used an identical architecture for all generative models (generator and discriminators depicted in Fig. 5.3) for fairness in comparison. Additionally, we have used the same settings for these networks.

Fig. 5.4 shows that the average number of learned modes has an increasing behavior of up to 20k iterations for MMD and μ-Fisher IPMs. For Wasserstein and Cramér IPMs, this behavior reaches around 26k iterations. Among these, the Sobolev IPM keeps its incremental behavior up to 30k iteration with considerable bias (along the y-axis). That demonstrates the higher performance of f_ϑ in capturing the local distribution of spectrograms in the restricted Sobolev space compared to other IPM. However, it does not immune our generative model against the

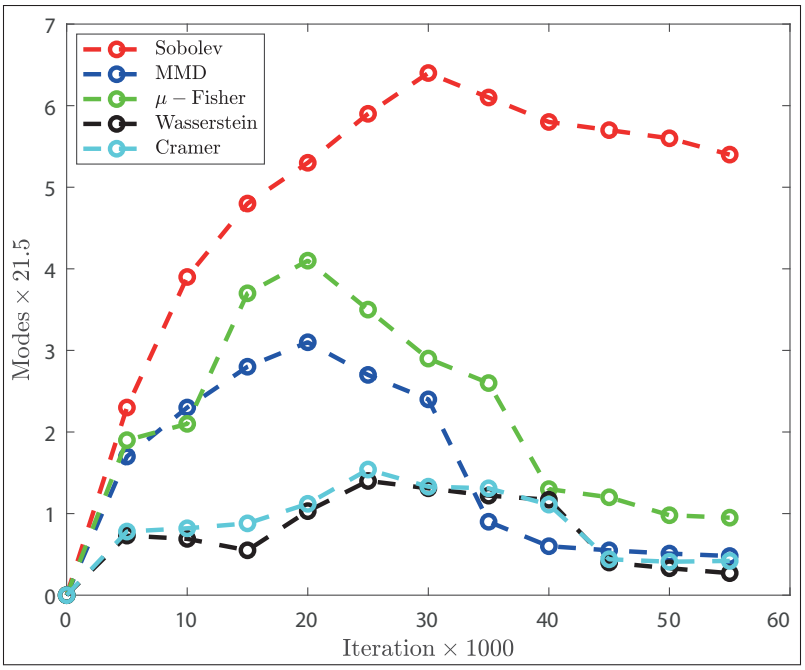

Figure 5.4 Monitoring the average learned modes (per batch size of 2×512) by our GAN model during training on SP_{STFT} with different IPMs indicates potential collapse over the total number of iterations

mode collapse issue. As depicted in Fig. 5.4, our GAN gradually starts losing sample modes after 31k iterations. For tackling this issue, we used OR, spectral normalization (Miyato *et al.*, 2018), and early stopped at checkpoints before the collapse.

Since there is a direct relationship between stability and generalizability of the GAN and our proposed defense algorithm, even a partially unstable generator network might result in absolute divergence in the chordal distance adjustment operation. In other words, if the GAN model is not comprehensive enough in terms of the number of learned modes, the process shown in Fig. 5.2 might never converge. This poses more concerns for long signals with too much environmental noise[45]. Additionally, for multi-speaker speech signals, our proposed Sobolev-DGANs not only might not be able to learn enough modes but also might recover adversarial perturbation after the i-STFT procedure. We believe that employing more constraining conditions on both

[45] Including reverberation and echo.

the generator and discriminators may improve model stability. Moreover, conditioning the discriminator networks aligned with time-distributed filter sets can provide more distinctive features for the discriminator network to resolve the multi-speaker issue. We are determined to address these issues in future work.

5.6 Conclusion

In this paper, we proposed a novel approach for defending speech-to-text models against end-to-end adversarial attacks. Our approach is based on reconstructing signals from synthesized spectrograms and approximated phase vectors. For spectrogram synthesis, we use a multi-discriminator GAN defined in the restricted Sobolev space. Our GAN generator network requires a safe input vector achievable through an iterative spectrogram subspace projection operation using the adjusted chordal distance. To improve our implemented generative model's performance, we impose a constraint for the critic function that learns discrepancies between real and synthesized sample distributions.

We evaluated our defense approach against six strong white and black-box adversarial attacks on advanced DeepSpeech, Kaldi, and Lingvo victim models. The proposed defense approach, averaged over the total number of experiments, outperformed other algorithms according to WER and SLA metrics. Furthermore, we used four objective quality metrics for measuring the impact of running defense algorithms on speech signals. For the majority of the cases, our defense approach demonstrated higher signal quality compared to other algorithms.

CONCLUSION AND RECOMMENDATIONS

There are numerous real-life applications for environmental sound classification and automatic speech transcription. For instance, context-aware computing systems use sound recognition algorithms for scene understanding. Likewise, voice command (including voice assistance systems such as Siri in Apple products) applications embedded into smartphones, modern TVs, autonomous vehicles, etc., employ built-in speech recognition models. All these systems are data-driven, and their performance is highly dependent on the power of their classification architectures in correctly capturing sample distributions and the volume of the training dataset. Over the last decade and particularly after the growth of deep learning algorithms, many strong unsupervised, semi-supervised, and supervised architectures have been introduced both for classification and data augmentation purposes. Nowadays, the recognition accuracy of the state-of-the-art data-driven ESC and ASR models is competitive to human-level of understanding. However, it has been proven that such advanced recognition systems are extremely vulnerable against several white and black-box adversarial attacks. This poses a major security concern against data-driven models and negatively affects their prediction accuracy. During the last few years, many defense algorithms have been introduced. Nevertheless, there is still no reliable approach for securing ESC and ASR models against strong adversarial attacks.

Unfortunately, there is no consensus on the definition of a reliable defense algorithm. Hence we explained our implications from reliability in the first chapter. As stated, a reliable defense approach should meet at least one of our reliability conditions to avoid offering a false sense of security against adversarial attacks.

1. It should prevent gradient obfuscation or the shattering of the Jacobian matrix;
2. It should make a reasonable trade-off among recognition accuracy, adversarial attack robustness, and computational complexity of the algorithm to work in real-time;
3. The classifier architecture should be designed to reflect a strong configuration that maximizes the cost of attack for the adversary;

4. Complying with each of these conditions should not conflict with another.

To the best of our knowledge, most of the defense approaches we reviewed in this thesis, at least partially, violate our reliability conditions, which motivated us to pursue our research toward developing reliable defense algorithms for ESC and ASR systems.

In Chapter 2, we proposed an ensemble-based classification architecture for ESC. This architecture employs a GAN in its back-end for augmenting DWT spectrograms and a random forest algorithm in its front-end for classification. This configuration helps to make a trade-off between recognition accuracy and model robustness against adversarial attacks in compliance with our defense reliability conditions. Our evaluations on four benchmark datasets, namely ESC-10, ESC-50, UrbanSound8K, and DCASE-2017, corroborated the superior performance of our classification approach over conventional and deep learning-based classifiers.

In Chapter 3, we developed a novel approach for securing ESC models from a large collection of targeted and non-targeted adversarial attacks. We argued that conventional classifiers such as SVMs are often more robust against adversarial attacks. On the other hand, they cannot usually outperform dense CNNs (in terms of model generalizability and prediction accuracy for non-adversarial signals). Therefore, to make an appropriate trade-off between attack robustness and prediction performance, we exploited the nonlinear SVM in the front-end and employed a CNN-based autoencoder in the back-end of our classification framework. Furthermore, we implemented logarithmic visualization, color compensation, highboost filtering, and dimensionality reduction operations in the back-end before the autoencoding block to expand the learning space for the front-end classifier. However, we do not run any of these operations directly on the test signal in runtime so as to meet all of our predefined defense reliability conditions. As it has been experimentally proven, our proposed algorithm outperforms other approaches both in terms of prediction accuracy and robustness against adversarial attacks.

In Chapter 4, we investigated the effect of spectrogram settings on victim classifier's recognition performance and fooling rate. More specifically, we identified major spectrogram settings which can considerably increase the cost of attack for the adversary (averaged over the allocated budgets). This investigation is primarily in line with our third defense reliability condition in developing inherently strong algorithms against adversarial attacks. As part of our extensive experiments, we experimentally proved that compared to DWT and STFT, the MFCC has a relatively lower adversarial transferability ratio among three advanced CNN architectures.

Lastly, in Chapter 5, we developed an implicit reactive defense approach against adversarial attacks for end-to-end transcription systems. Our proposed algorithm is synthesis-based and it complies with all defense reliability conditions mentioned above. The generative model used for signal synthesis is a multi-discriminator GAN conditioned in the restricted Sobolev space. We used the Sobolev integral probability metric for training the critic function since it is closely related to STFT-based representations. To avoid unnecessary complications, we implemented simple yet effective architectures for both the generator and discriminator networks. Additionally, we characterized a constraining technique for improving the stability of our GAN in adverse environmental scenarios. We evaluated this defense approach against the strong white and black-box adversarial attacks benchmarked over the cutting-edge speech-to-text transcription systems, namely DeepSpeech, Kaldi, and Lingvo. Our experiments corroborate the superior performance of our proposed defense algorithm compared to other state-of-the-art approaches.

Thus far in this thesis, we introduced four defense algorithms towards addressing the security issues of ESC and ASR systems against a variety of targeted and non-targeted as well as white and black-box adversarial attacks. Moreover, we mentioned the limitations of our works in every chapter. In the following subsection, we also summarize a few of such limitations and major challenges which determined us to tackle them in our future works.

Future Works

The overall findings in this thesis suggest the following directions for future works in the context of securing ESC and ASR systems:

1. *Optimizing the trade-off between recognition accuracy and model's robustness.* This thesis proposed three defense techniques mainly for making a trade-off between a model's recognition performance and its robustness against adversarial attacks (i.e., Chapter 2, 3, and 4). According to our conducted experiments on the benchmarking datasets, these algorithms can make almost a solid trade-off in this regard. However, we believe that employing more strict regularization schemes (mainly for the front-end classifier) can better optimize the achieved trade-offs.
2. *Improving the stability of GANs for synthesis-based (1D signal and representation-level) defense approaches.* We introduced three signal synthesis-based defense algorithms in this thesis (i.e., Chapter 5, Appendix II, and Appendix IV). Although we used different IPMs, critic functions, learning spaces (e.g., Lebesgue, Sobolev, etc.), regularizations, and architectures for training these GANs, we still noticed signs of instability and mode collapse issues at larger iterations. For addressing these critical issues, we developed a novel conditioning trick (Esmaeilpour, Sallo, St-Georges, Cardinal & Koerich, 2020d)[46], however we could only delay collapse onsets. Since achieving a more stable GAN contributes substantially to a more solid synthesis-based defense approach, there is a constant need to devise stricter conditioning tricks to avoid extreme mode collapse issues.
3. *Preserving the quality of signals during the defense procedure.* Unfortunately, both explicit and implicit reactive defense approaches negatively affect the quality of the input signals. This potential side-effect has been briefly mentioned in Chapter 5 and we used a few signal quality metrics for evaluating the performance of the defense algorithms in terms of quality preservation. Degrading the quality of the signals after filtration operation by a

[46] This paper is not included in this thesis.

defense algorithm poses a major concern since this operation might bypass the potential adversarial perturbation, However it might also make the filtered signal uninterpretable (impossible to transcribe) to the transcription systems. Our recommendation for tackling this concern is to employ a psychoacoustic loss function (similar to the approach introduced by Szurley & Kolter (2019)) for the defense algorithm.

4. *Developing a faster adversarial attack algorithm.* As mentioned in the first chapter, adversarially training is among the reliable defense techniques since it does not obfuscate gradient vectors. However, it requires developing a very fast adversarial attack to make the adversarially training procedure feasible. In response to this concern, we developed a faster attack formulation (see Appendix V), Nevertheless it might fade out the adversarial perturbation after a few playbacks over the air. Our initial experiments show that incorporating restricted probability metrics such as MMD (Dziugaite *et al.*, 2015) into the attack optimization formulation might improve the resiliency of the perturbation over consecutive playbacks.

LIST OF PUBLICATIONS

Journal Articles

1. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2019). "Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network." Elsevier Applied Soft Computing, 86, 105912.
2. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2019). "A robust approach for securing audio classification against adversarial attacks." IEEE Transactions on Information Forensics and Security (TIFS), 15, 2147-2159.
3. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2020). "From sound representation to model robustness." Currently under review at Elsevier Applied Acoustics Journal.
4. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2021). "Multi-Discriminator Sobolev Defense-GAN Against Adversarial Attacks for End-to-End Speech Systems." Currently under review at IEEE Transactions on Information Forensics and Security (TIFS) Journal.
5. M. Esmaeilpour, P. Cardinal and A. L. Koerich, "Cyclic Defense GAN Against Speech Adversarial Attacks," in IEEE Signal Processing Letters, vol. 28, pp. 1769-1773, 2021, doi: 10.1109/LSP.2021.3106239.

Conference Papers

1. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2022). "Towards Robust Speech-to-Text Adversarial Attack". Currently under review at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2022).
2. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2021). "Class-Conditional Defense GAN Against End-to-End Speech Attacks." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2021), pp. 2565-2569.
3. Sallo, R. A., Esmaeilpour, M., and Cardinal, P. (2020-2021). "Adversarially Training for Audio Classifiers." In 25th International Conference on Pattern Recognition (ICPR), (pp. 9569-9576). IEEE.

4. Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2020). "Detection of adversarial attacks and characterization of adversarial subspace." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2020), pp. 3097-3101.

APPENDIX I

DETECTION OF ADVERSARIAL ATTACKS AND CHARACTERIZATION OF ADVERSARIAL SUBSPACE

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in « 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) » in 2020.

Abstract

Adversarial attacks have always been a serious threat for any data-driven model. In this paper, we explore subspaces of adversarial examples in unitary vector domain, and we propose a novel detector for defending our models trained for environmental sound classification. We measure chordal distance between legitimate and malicious representation of sounds in unitary space of generalized Schur decomposition and show that their manifolds lie far from each other. Our front-end detector is a regularized logistic regression which discriminates eigenvalues of legitimate and adversarial spectrograms. The experimental results on three benchmarking datasets of environmental sounds represented by spectrograms reveal high detection rate of the proposed detector for eight types of adversarial attacks and it also outperforms other detection approaches.

2. Introduction

In the field of sound and speech processing, it is very common to use 2D representations of audio signals for training data-driven algorithms. Such 2D representations have lower dimensionality than audio waveforms and they easily fit advanced deep learning architectures mainly developed for computer vision applications. Mel frequency cepstral coefficient (MFCC), short-time Fourier transform (STFT), discrete wavelet transform (DWT) are among the most

pervasive 2D representations which essentially visualize frequency-magnitude distribution of a given reconstructed signal over time. Thus far, the best sound classification accuracy has been achieved for deep learning algorithms trained on 2D representations (Boddapati *et al.*, 2017; Esmaeilpour *et al.*, 2020a). However, it has been shown that despite achieving high performance, the approaches based on 2D representations are very vulnerable against adversarial attacks (Esmaeilpour *et al.*, 2020; Koerich, Esmailpour, Abdoli, Jr. & Koerich, 2019). Unfortunately, this poses a security issue because crafted adversarial examples not only mislead the target model toward a wrong label, but are also transferable to other models such as support vector machines (SVM) (Esmaeilpour *et al.*, 2020).

There are some discussions about existence, origin, and behavior of adversarial examples, notably their linear characteristics (Goodfellow *et al.*, 2015), but there is no reliable approach to discriminate their underlying subspace(s) compared to legitimate examples. In an effort to characterize possible adversarial subspace detectors based on a statistical comparison on predictions of the victim model have been introduced. Feinman et al. (Feinman, Curtin, Shintre & Gardner, 2017) have proposed to estimate kernel density (KD) and Bayesian uncertainty (BU) of a deep neural network (DNN) for triplets of legitimate, noisy, and adversarial examples. All these measurements have been carried out with the assumption of approximating a DNN to a deep Gaussian process and they result in high ratios of KD and BU for adversarial examples compared to legitimate and noisy samples. Measuring maximum mean discrepancy and energy distance of examples are two other statistical metrics for investigating adversarial manifolds using divergence of model predictions for clusters of datapoints (Grosse, Manoharan, Papernot, Backes & McDaniel, 2017). In addition to these output-level statistical measurements, logits of adversariality have been carefully assessed in each subnetwork placed on top of some hidden units of the victim model (Metzen, Genewein, Fischer & Bischoff, 2017) as well as measuring instability of potential layers to perturbations (Rouhani, Samragh, Javidi & Koushanfar, 2017). Ma et al. (Ma *et al.*, 2018) presented a comprehensive study for characterizing adversarial manifolds and introduced local intrinsic dimensionality (LID) score, which measures ℓ_2 distance of network prediction for a given example compared to prediction logits of its k neighbours at

each hidden unit. The actual detector is a logistic regression binary classifier trained on one class made up of LID vectors of legitimate and noisy examples because they lie in a very close subspace and another class made up of LID vectors of adversarial examples generated by strong attacks. Experimental results on several datasets have shown the competitive performance of the LID detector compared to KD and BU (Ma *et al.*, 2018). Unfortunately, it has been shown that these adversarial detectors might fail to detect strong adversarial attacks in adverse scenarios (Carlini & Wagner, 2017a; Athalye *et al.*, 2018b), due to the difficulty in tuning detectors or even due to the particular characteristics of the datasets.

In this paper we show that adversarial manifolds lie far from legitimate and noisy spectrograms for short audio signals using a unitary space-based chordal distance metric. We also provide an algorithm to proactively detect potential malicious examples using generalized Schur decomposition (a.k.a. QZ decomposition) (Van Loan & Golub, 1983). This paper is organized as follows. Section 3 presents a brief explanation of unitary space of QZ and our adversarial detection algorithm. Experimental results on DWT representation of three environmental sound datasets are discussed in the Section 4. Conclusions and perspectives of future work are presented in the last section.

3. Adversarial Detection

Computing norm metrics is a common approach for measuring the similarity between crafted adversarial examples and their legitimate counterparts. In addition to basic norms such as l_2 and l_∞, human visual inference oriented metric has been also embedded in general optimization problems (Rozsa, Rudd & Boult, 2016). These similarity constraints are probably the most valuable clues in studying possible subspaces of crafted examples.

It has been shown that regardless of the category or type of adversarial attack, the generated examples, subject to a similarity constraint, lie in a sub-Cartesian space further than the legitimate ones (Ma *et al.*, 2018). However, this is tricky and may not work correctly for strong attacks (Athalye *et al.*, 2018b). Our detailed study of the failure cases of such detectors uncovered

imperfection of Cartesian metric space (distance-based) for exploring adversarial subspaces. Therefore, vector spaces that may discriminate between adversarial and legitimate manifolds can be very useful to build robust adversarial example detectors.

We investigate the mapping of input samples to the vector space of generalized Schur decomposition and the use of chordal distance to identify their underlying subspaces.

3.1 Schur Decomposition and Chordal Distance

For computing generalized Schur decomposition of two spectrograms denoted as M_1 and M_2 in a complex set $\mathbb{C}^{n\times n}$ there should exist unitary matrices Q and Z such that:

$$Q^H M_1 Z = T, \quad Q^H M_2 Z = S \tag{A I-1}$$

where S and T are upper (quasi) triangular and Q^H denotes the conjugate transpose of Q. The eigenvalues (λ) of M_1 and M_2 can be approximated as:

$$\lambda(M_1, M_2) = \{t_{ii}/s_{ii} : s_{ii} \neq 0\} \tag{A I-2}$$

where t_{ii} and s_{ii} are diagonal elements of T and S, respectively, and $\lambda(M_1, M_2) = \mathbb{C}$ for some zero-valued diagonal entries of S and T. In other words, super-resolution similarity between two spectrograms can be calculated as:

$$\det(M_1 - \lambda M_2) = \det\left(QZ^H\right) \prod_{i=1}^{n} (t_{ii} - \lambda s_{ii}) \tag{A I-3}$$

According to the Bolzano-Weierstrass theorem (Van Loan & Golub, 1983), the bounded basis matrices $\left\{(Q_k, Z_k)\right\}$ which are definite seris support $\lim_{i\to\infty}(Q_{k_i}, Z_{k_i}) = (Q, Z)$. The unitary subsequence of Z_k leads to:

$$Z_k^H (M_{2,k}^{-1} Q_k) = S_k^{-1} \tag{A I-4}$$

which asymptotically implies $Q_k^H \left(M_1 M_2^{-1} Q_k \right)$ equivalent to generic Schur decomposition of $M_1 M_{2,k}^{-1}$ for nonsingular basis matrices of $\left\{ M_{2,k} \right\}$.

In perturbing the spectrogram M_i where $\widetilde{M_i} \simeq M_i + \epsilon$ increases considerably the chance of noticeable variations in the resulting eigenvalues/eigenvectors (Van Loan & Golub, 1983). Theoretically, we can measure it using the chordal metric where the pencil of $\vec{\mu}_i M_i - \widetilde{M_i}$ is the point of interest for $\mu_i \in \left\{ \left(t_{ii}, s_{ii} \right) | s_{ii}/t_{ii} \right\}$ perturbed by ϵ as conditioned in Eq. A I-5.

$$\left\| M_i - \widetilde{M_i} \right\|_2 \simeq \epsilon_i \tag{A I-5}$$

where ϵ_i is a very small perturbation. The chordal distance for the vectors of eigenvalues associated with pencil of $\vec{\mu}_i M_i - \widetilde{M_i}$ can be measured by Eq. A I-6 (Van Loan & Golub, 1983).

$$\text{chord}(\lambda_i, \lambda_{i,\epsilon}) = \frac{\left| \lambda_i - \lambda_{i,\epsilon} \right|}{\sqrt{1 + \lambda_i^2} \sqrt{1 + \lambda_{i,\epsilon}^2}} \tag{A I-6}$$

where pencils are neither necessarily bound to be normalized nor differentiable. For any adversarial attack that perturbs a legitimate spectrogram M_i by ϵ_i, we compute chordal distance as Eq. A I-6 and we compare the distances obtained to find separable manifolds for legitimate and adversarial examples.

3.2 Adversarial Subspace

We can explore the properties of adversarial examples using chordal distance in unitary space of eigenvectors where each spectrogram is represented by basis functions Q_i and Z_i. For any legitimate and adversarial spectrograms, the chordal distance between their associated eigenvalues $(\lambda, \lambda_{i,\epsilon})$ must satisfy the constraint defined in Eq. A I-7 (Van Loan & Golub, 1983).

$$\text{chord}(\lambda_i, \lambda_{i,\epsilon}) \leq \frac{\epsilon}{\sqrt{\left[\left(y^H M_i x \right) + \left(y^H \widetilde{M_i} x \right) \right]^2}} \tag{A I-7}$$

where x and y satisfy $M_i x = \lambda \widetilde{M_i} x$ and $y^H M_i = \lambda y^H \widetilde{M_i}$ for the symmetric in the upper bound of M_i and $\widetilde{M_i}$. The extreme case for the defined pencil may happen when both s_{ii} and t_{ii} are zero. Therefore, we can replace their division with a small random value close to their neighbours.

Not only satisfying Eq. A I-5 is required for properly computing chordal distance of eigenvalues, but it must also be part of the optimization procedure of any adversarial attack because the perturbation value ϵ should not be perceivable. For adversarial perturbations, an adjustment of the chordal distance by a factor γ is also required (chord$(\lambda_i, \lambda_{i,\epsilon}) + \gamma_i$). The value of such a hyperparameter should be very small and associated to mean eigenvalue, otherwise it might cause ill-conditioning cases. We examine the effects of different pencil perturbations on the chordal distance and inequality of Eq. A I-7 from random noisy to carefully optimized adversarials in Section 4.

3.3 Adversarial Discrimination

In practice, detecting adversarial examples using chordal distance for a given input requires access to its reference spectrogram as well as to the perturbation ϵ. However, this is not feasible for real life applications. To circumvent this issue, we compare eigenvalues of legitimate and adversarial examples to draw a decision boundary between them. To this end, we train a logistic regression on the eigenvalues of legitimate and adversarial examples as shown in Algorithm I-1.

For every spectrogram pairs randomly picked from an identical class, we compute their eigenvalues using QZ decomposition. We assume that spectrograms have been generated for short audio signals and they share significant similarities, especially when they are split into smaller batches.

For a test spectrogram M, its eigenvalues generated by Schur decomposition are used as arguments for the front-end classifier (detector) as relations of these two decompositions have been explained in Section 3.1. Generalizing this algorithm to a multiclass problem requires computing eigenvectors of inter-class samples sharing no significant similarity due to causing ill-conditioned decomposition for pencils.

Algorithm-A I-1 Discriminating adversarial examples from legitimate ones using their associated eigenvectors

Algorithm-A: Discriminating adversarial signals from original representations.
Input: C_{leg}: Class of legitimate samples
Output: Detector$\left[\text{schur}(M)\right]$ for the given test spectrogram M

$\Lambda_{leg} = [] \, , \Lambda_{adv} = [];$ /* lists */
for *all* $B_{leg} \in C_{leg}$; /* B_{leg}: legitimate batch */
do
  B_{adv} := adversarial attack on B_{leg}; /* B_{adv}: adversarial batch */
  $\overrightarrow{\lambda}_{leg} = \text{eigen}\left[\text{qz}\left(B_{leg}\left[i\right], B_{leg}\left[j\right]\right)\right];$ /* $i \neq j$ */
  $\overrightarrow{\lambda}_{adv} = \text{eigen}\left[\text{qz}\left(B_{adv}\left[i\right], B_{adv}\left[j\right]\right)\right];$ /* $i \neq j$ */
  Λ_{leg}.append $\left(\overrightarrow{\lambda}_{leg}\right)$, Λ_{adv}.append $\left(\overrightarrow{\lambda}_{adv}\right)$
end for
Detector$\left[\text{schur}(M)\right]$ = train a classifier on (Λ_{leg}, Λ_{adv});

4. Experimental Results

We have evaluated the performance of computing chordal distance on adversarial detection and the performance of the proposed detector in adverse scenarios on three environmental sound datasets: ESC-10, ESC-50, and UrbanSound8k (Piczak, 2015b; Salamon *et al.*, 2014a). The first dataset includes 400 five-second length audio recordings of 10 classes. It is a simplified version of ESC-50 which has 2000 samples of 50 classes. The UrbanSound8k dataset contains 8732 samples ($\leq$ 4s) of 10 classes and it provides more diversity both in terms of quality and quantity than the first two datasets.

We apply pitch-shifting operation as part of 1D signal augmentation as proposed in (Esmaeilpour *et al.*, 2020a). This low-level data augmentation increases the chance of learning more discriminant features by the classifier, especially for ESC-10 and ESC-50 compared to UrbanSound8k. Four pitch-shifting scales $(0.75, 0.9, 1.15, 1.5)$ are applied to each sample to add four new samples to the legitimate sets. These hyperparameter values are the most effective for these datasets (Esmaeilpour *et al.*, 2020a). The wavelet mother function used for producing DWT spectrogram representations is complex Morlet. Sampling frequency and frame length are set to

8kHz and 50ms for ESC-10 and UrbanSound8k and 16kHz and 30ms for ESC-50 with fixed overlapping ratio of 0.5 for all datasets (Boddapati *et al.*, 2017). The convolution of the Morlet function with the signal produces a complex function with considerable overlap between real and imaginary parts. Therefore, for representing real spectrograms we use linear, logarithmic, and logarithmic real visualizations. The first visualization scheme highlights high-frequency magnitudes which denote high variation areas. Low-frequency information is characterized by a logarithmic operation which expands their distances. Energy of the signal, which is associated with the signal's mean, is obtained by applying a logarithmic filter on the real part. Since the frequency-magnitude of a signal distributed over time has variational dimensions, none of the three visualizations produce square spectrograms. Hence, we bilinearly interpolate each spectrogram to fit square size with respect to this constraint of QZ decomposition. The actual size of the spectrograms for ESC-10 and ESC-50 is 1536×768 and 1168×864 for UrbanSound8k because the latter has shorter audio recordings of at least one second. Final size of spectrograms after downsampling and interpolation is 768×768. This lossy operation may remove some pivotal frequency information and consequently it may decrease the performance of the classifier. However, obtaining the highest recognition accuracy is not our point of interest in this paper, but studying adversarial subspaces.

We use an SVM and a convolutional neural network (CNN) as victim classifiers, to evaluate the detection rate of the proposed detector for a variety of adversarial attacks. In SVM configuration, we use scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot & Duchesnay, 2011a) with a grid search. Linear, polynomial, and RBF kernels have been evaluated on the 2/3 of the shuffled datasets (training and development). The best recognition accuracy on the test set was achieved with the RBF kernel with about 72.056%, 71.257%, 72.362% for ESC-10, ESC-50 and UrbanSound8k datasets, respectively. The proposed CNN has four convolutional layers with receptive field 3×3, stride 1×1, and 128, 256, 512, and 128 filters, respectively. On top of the last convolution layer there are two fully connected layers of sizes 256 and 128. All layers use ReLU activation function, except the output layer for which softmax is used. Batch and weight

Table 1.1 The mean γ values for justifying chordal distances of adversarial examples, the corresponding mean perturbation and the recognition accuracy of victim models (CNN & SVM) on adversarial sets

	FGSM	BIM-a	BIM-b	JSMA	CWA	Opt	EA	LFA
γ	7 ± 0.12	6 ± 0.03	8 ± 0.01	7 ± 0.17	11 ± 0.09	10 ± 0.27	8 ± 0.39	12 ± 0.16
ℓ_2	5.637	4.015	6.371	6.187	4.426	5.067	NA	NA
Accuracy (AUC score %)	3.036	6.017	4.964	3.189	6.237	8.143	15.157	17.845

NA: Not Applicable.

normalization have been applied at all convolutional layers. Such a CNN can achieve recognition performance of 73.415%, 73.674%, and 75.376% for ESC-10, ESC-50, and UrbanSound8k datasets respectively on the 1/3 test set.

We attack the CNN by fast gradient sign method (FGSM) (Goodfellow *et al.*, 2015), basic iterative methods (BIM-a and BIM-b) (Kurakin *et al.*, 2016), Jacobian-based salience map attack (JSMA) (Papernot *et al.*, 2016d), optimization-based attack (Opt) (Liu *et al.*, 2016), and Carlini& Wagner attack (CWA) (Carlini & Wagner, 2017b). For the SVM model, we use label flipping attack (LFA) (Xiao *et al.*, 2012) and evasion attack (EA) (Biggio *et al.*, 2013). Overall, for each legitimate DWT spectrogram (M_i), eight adversarial examples are crafted ($\widetilde{M}_{i,j}$ for $j = 1 \ldots 8$). For each pencil of $\mu_i M_i - \widetilde{M}_{i,j}$, we measure their chordal distances using Eq. A I-6, then for a random unit 2-norm x and y matrices, we check for the inequality of Eq. A I-7 and required γ adjustments. Similarly, we add random Gaussian noise to each M_i with zero mean and $\sigma \in \left\{0.01, 0.02, 0.04, 0.05\right\}$ and build pencil of $\mu_i M_i - N_{i,k}$ where $N_{i,j}$ for $k = 1 \ldots 4$ denote the noisy spectrograms which also satisfy Eq. A I-5. Table A-1.1 summarizes the adjustment of γ required for crafted adversarial examples to satisfy Eq. A I-7. For the generated noisy examples, an adjustment of γ to 0.5 ± 0.012 is needed, which is averaged over different values of σ. Considerable displacement between chordal distance adjustments required for adversarial and noisy spectrogram sets denote their non-identical and dissimilar subspaces.

For evaluating the performance of the Algorithm I-1 in discriminating adversarial from legitimate examples, we use all the attacks mentioned above for crafting B_{adv}. Regularized logistic regression has been used as the front-end classifier for discriminating Λ_{leg} from Λ_{adv}. We compare the performance of the proposed detector with LID, KD, BU, and the combination

Table 1.2 Mean class-wise comparison of the AUC (%) achieved by the adversarial detectors for spectrograms attacked with eight adversarial attacks. The best results are highlighted in bold.

Detector	CNN						SVM	
	FGSM	BIM-a	BIM-b	JSMA	CWA	Opt	EA	LFA
KD	65.234	88.097	87.914	63.552	61.025	86.105	55.479	63.659
BU	39.025	80.673	55.474	80.603	58.022	69.207	57.861	67.610
KD+BU (Using both)	74.381	91.154	88.243	89.251	64.349	90.461	58.330	69.008
LID (averaged over its k neighbours)	79.299	93.097	94.671	91.665	75.297	**94.781**	70.981	**71.239**
Proposed	**84.132**	**96.519**	**95.349**	**94.375**	**89.957**	93.309	**75.227**	71.198

KD+BU. Table A-1.2 shows that the proposed detector outperforms other detectors for the majority of the attacks. The proposed detector can be used with MFCC and STFT representations or even other datasets commonly used for computer vision applications. The key challenge in this detector is its sensitivity to intra-class sample similarities, otherwise it may not satisfy Eq. A I-7, especially for black-box multiclass discrimination. Moreover, the performance of the front-end classifier is dependent on the volume of the training eigenvectors and their similarities. Providing large enough short audio signals with high intra similarity considerably increases the chance of finding a comprehensive decision boundary among dissimilar eigenvectors.

5. Conclusion

Since adversarial examples are visually very similar to the legitimate samples, differentiating their underlying subspaces is very challenging in Cartesian metric space. In this paper we show that the offset between subspace of legitimate spectrograms compared to their associated adversarial examples can be measured by chordal distance defined in unitary vector space of generalized Schur decomposition. Using this metric, we demonstrated that manifold of adversarial examples lie far from legitimates and noisy samples which have been slightly perturbed by Gaussian noise.

In order to detect any adversarial attack when there is no access neither to reference spectrogram nor adversarial perturbation, we proposed a detector which is a regularized logistic regression model for discriminating eigenvalues of malicious spectrograms from legitimate ones. Experimental results on three environmental sound datasets have shown that the proposed detector outperforms other detectors for six out of eight different adversarial attacks.

For future studies, we would like to improve chordal distance to better characterize adversarial manifolds and also study possibility of encoding this metric directly into the adversarial detector.

APPENDIX II

ADVERSARIALLY TRAINING FOR AUDIO CLASSIFIERS

Raymel Alfonso Sallo[a] , Mohammad Esmaeilpour[a] , Patrick Cardinal[a]

⋄ *The first and second authors made equal contributions.* ⋄
[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in « 25th IEEE International Conference on Pattern Recognition (ICPR) » in 2020-2021.

Abstract

In this paper, we investigate the potential effect of the adversarially training on the robustness of six advanced deep neural networks against a variety of targeted and non-targeted adversarial attacks. We firstly show that the ResNet-56 model trained on the 2D representation of the discrete wavelet transform appended with the tonnetz chromagram outperforms other models in terms of recognition accuracy. Then we demonstrate the positive impact of adversarially training on this model as well as other deep architectures against six types of attack algorithms (white and black-box) with the cost of the reduced recognition accuracy and limited adversarial perturbation. We run our experiments on two benchmarking environmental sound datasets and show that without any imposed limitations on the budget allocations for the adversary, the fooling rate of the adversarially trained models can exceed 90%. In other words, adversarial attacks exist in any scales, but they might require higher adversarial perturbations compared to non-adversarially trained models.

2. Introduction

The existence of adversarial attacks has been characterized for data-driven audio and speech recognition models for both waveform and representation domains (Carlini & Wagner, 2018; Esmaeilpour *et al.*, 2020). During the past years, many strong white and black-box adversarial

algorithms have been introduced which they basically recast costly optimization problems against victim classifiers. Unfortunately, these attacks effectively degrade the classification performance of almost all data-driven models from conventional classifiers (e.g., support vector machines) to the state-of-the-art deep neural networks (Esmaeilpour *et al.*, 2020b). This poses an extreme growing concern about the security and the reliability of the classifiers.

The typical approach in crafting adversarial examples is to solve an optimization problem in order to obtain the smallest possible perturbations for the legitimate samples, undetectable by a human, aiming at fooling the classifier. The commonly used measures to compare the altered sample with the original one are l_2 or l_∞ similarity metrics. The computational complexity of this optimization process is dependent on the dimensions of the given input samples. Consequently, it requires considerable computational overhead for high dimensional data, even in the case of short audio signals (Carlini & Wagner, 2018). However, regardless of the computational cost of the attacks, this threat actively exists for any end-to-end audio and speech classifier. Since the highest recognition accuracies have been reported on 2D representations of audio signals (Esmaeilpour *et al.*, 2020; Boddapati *et al.*, 2017), the optimized attack algorithms developed for computer vision applications such as fast gradient sign method (FGSM) (Goodfellow *et al.*, 2015) led to security concerns for audio classifiers (Esmaeilpour *et al.*, 2020b).

Although some approaches have been introduced for defending victim models against adversarial attacks, there is not yet a reliable framework achieving the required efficiency. Based on the detailed discussion in (Athalye *et al.*, 2018b), common defence algorithms usually obfuscate gradient information but running stronger attack algorithms against them consistently fool these detectors. Unfortunately, vulnerability against adversarial attacks is an open problem in data-driven classification and though the generated fake examples look very similar to noisy samples, they lie in dissimilar subspaces[47]. (Esmaeilpour *et al.*, 2020b; Ma *et al.*, 2018). It has been shown that adversarial examples lie in the manifolds marginally over the decision boundary

[47] Technically, a subspace is a vector space with some specific properties such as nonlinearity (Strang, 1993; Van Loan & Golub, 1983). Subspaces are definable by both Cartesian and non-Cartesian coordinates and can represent a structure in any n-dimensional spaces (Van Loan & Golub, 1983). More information are available in (Watkins, 2007).

of the victim classifier, where the model lacks of generalizability (Esmaeilpour *et al.*, 2020b). Therefore, integrating these examples into the training set of the victim classifier could improve the robustness. This approach, known as adversarially training (Goodfellow *et al.*, 2015), might be a more reasonable defense approach without shattering gradient vectors (Athalye *et al.*, 2018b). However, there is no guarantee for the safety of the adversarially trained classifiers (Tramèr *et al.*, 2017).

Although there are some discussions in the computer vision domain about the negative effect of adversarially training on the recognition performance of the victim classifiers (Papernot *et al.*, 2016c), to the best of our knowledge, this potential side effect has not been yet studied for the 2D representation of audio signals. We address this issue in this paper and report our results on two benchmarking environmental sound datasets. Specifically, our main contributions in this paper are:

1. characterizing the adversarially training impact on six advanced deep neural network architectures for diverse audio representations[48],
2. demonstrating that deep neural networks especially those with residual blocks have higher recognition performance on tonnetz features concatenated with DWT spectrograms compared to STFT representations,
3. showing the adversarially trained AlexNet model outperforms ResNets with limiting the perturbation magnitude,
4. experimentally proving that although adversarially training reduces recognition accuracy of the victim model, it makes the attack more costly for the adversary in terms of required perturbation.

The rest of this paper is organized as follows. In Section 3, we review some related works about adversarial attacks developed for 2D domains. Details about signal transformation and 2D representation production are provided in Section 4 and 5, respectively. In Section 6, we briefly introduce our selected front-end audio classifiers which are state-of-the-art deep learning architectures. The adversarial attack procedures and budget allocation for the adversary are

[48] Namely, MFCC, STFT, and DWT.

discussed in Section 7. Accordingly, section 8 explains the adversarially training framework and obtained results are summarized in Section 9.

3. Related Works

There is a large volume of published studies on attacking classifiers using different optimization techniques aiming to effectively disrupt their recognition performances. In this paper, we focus on five strong white-box targeted and non-targeted attack algorithms which have been reported to be very destructive when used on advanced deep learning models trained on audio representations (Esmaeilpour *et al.*, 2020). Moreover, we also use a black-box adversarial attack, based on the gradient approximation, against the victim classifiers .

The fast gradient sign method is a well-established baseline in targeted adversarial attack. The computational cost of this one-shot approach at runtime is low, taking advantage of the linear characteristics in deep neural networks. Kurakin *et al.* (Kurakin *et al.*, 2016) introduced an iterative version of FGSM, known as the basic iterative method (BIM), for running stronger attacks against victim classifiers and is formulated at:

$$\mathbf{x'}_{n+1} = \text{clip}\left\{\mathbf{x'}_n + \zeta \text{sgn}\left(\nabla_{\mathbf{x}} J\left[\mathbf{x'}_n, l(\mathbf{x})\right]\right)\right\} \tag{A II-1}$$

where the legitimate and its associated adversarial examples are represented by $\mathbf{x}$ and $\mathbf{x'}$, respectively. The initial state in this recursive formulation is $\mathbf{x'}_0 = \mathbf{x}$ in the ϵ-neighbourhood (the distance measured by a similarity metric such as l_2) of the legitimate manifold. This is followed by a clipping operation for keeping the adversarial perturbation within $[-\epsilon, \epsilon]$. Moreover, $l(\mathbf{x})$ and $\text{sgn}(\cdot)$ stand for the label of $\mathbf{x}$ and the general sign function. In Eq. A II-1, the step size $\zeta = 1$, though it is tunable according to the adversary's wishes. Two types of optimizations can be used with Eq. A II-1: (1) optimizing up to reach the first adversarial example (BIM-a) and (2) continuing the optimization up to a predefined number of iterations (BIM-b). For measuring the ϵ, two similarity metrics are suggested: l_∞ and l_2. In this work, we focus on the latter.

Gradient information of a deep neural network contains direction of intensity variation associated with the model decision boundary. Exploiting these information vectors for finding the least likely probability distribution is the key idea of the Jacobian-based Saliency map attack (JSMA) (Papernot *et al.*, 2016d). For the adversarial label l', this iterative attack algorithm runs against the model f and strives to achieve $l' = \arg\max_j f_j(\mathbf{x})$. The JSMA increases the probability of the target label l' while minimizing those of the other classes including the ground-truth using a saliency map as shown in Eq. A II-2.

$$S(\mathbf{x}, l')[i] = \left|J_{i,l'}(\mathbf{x})\right| \left(\sum_{j \neq l'} J_{i,j}(\mathbf{x})\right) \tag{A II-2}$$

where $J_{i,j}$ denotes the forward derivative of the model for the feature i computed as:

$$J_f[i, j](\mathbf{x}) = \frac{\partial f_j(\mathbf{x})}{\partial \mathbf{x}_i} \tag{A II-3}$$

the Jacobian vectors associated with label l' and values of the saliency map less or greater than zero (no variation shield), $S(\mathbf{x}, l')[i] = 0$. This white-box attack algorithm searches, iteratively, the feature index on which the perturbation will be applied in order to fool the model toward the target label l' using the similarity metric l_0.

The perturbation required for pushing a sample over the decision boundary of the victim classifier should be as minimal as possible. In a white box scenario, the optimization process uses local properties of the decision boundary. It has been shown that linearizing the boundary in the subspace of the original samples can yield to adversarial perturbation smaller than FGSM attack. This approach, known as the DeepFool attack, is shown in Eq. A II-4 (Moosavi-Dezfooli *et al.*, 2016):

$$\arg\min \|\epsilon\|_2, \quad \epsilon = -f(\mathbf{x})\mathbf{w}/\|\mathbf{w}\|_2^2 \tag{A II-4}$$

where the $\mathbf{w}$ refers to the weight function of the recognition model. Unlike other abovementioned adversarial attacks, DeepFool is a non-targeted attack and it iterates as many times as needed for pushing random samples to be marginally over the locally linearized decision boundary with the

condition of maximizing the prediction likelihood toward any labels other than the ground-truth. Though both l_∞ or l_2 measurement metrics can be used in the DeepFool attack, we use the latter in accordance with BIM algorithms.

Presumably, a straightforward approach for keeping an adversarial perturbation undetectable can be achieved by reducing its magnitude and distributing it over all input features. Additionally, not every feature should be perturbed and their gradient vectors should not be shattered. Following these two assumptions, Carlini and Wagner attack (CWA) has been introduced (Carlini & Wagner, 2017b). The general framework of their proposed algorithm is based on the following minimization problem:

$$\min \|\mathbf{x}' - \mathbf{x}\|_2^2 + c \cdot \mathcal{L}(\mathbf{x}') \tag{A II-5}$$

where the constant c is obtainable through a binary search. Finding the most appropriate value for this hyperparameter is very challenging since it may easily dominate the distance function and push the sample too far away from the adversarial subspace. Although in Eq. A II-5 the l_2 similarity metric for computing the adversarial perturbation ϵ is employed, CWA properly generalizes for both l_0 and l_∞. In the configuration of this adversarial attack, the loss function $\mathcal{L}$ is defined over the logits of Z for the trained model f as shown in the following equation:

$$\mathcal{L}(\mathbf{x}') = \max \left[\max_{i \neq l'} \{ Z(\mathbf{x}')_i - Z(\mathbf{x}')_{l'}, -\kappa \} \right] \tag{A II-6}$$

where κ controls the effectiveness and the adjacency of the adversarial examples to the decision boundary of the model. In this regard, higher values for this parameter in conjunction with a minimum ϵ-neighbourhood results in adversarial examples with higher confidence.

For achieving the overall unrestricted adversarial perturbation ($\|\epsilon\|_2$) with small enough magnitude, CWA solves Eq. A II-5 through the following optimization framework:

$$\min_{\rho} \left\| \frac{1}{2} \left(\tanh(\rho) + 1 \right) - \mathbf{x} \right\|_2^2 + c \cdot \mathcal{L} \left(\frac{1}{2} \tanh(\rho) + 1 \right) \tag{A II-7}$$

where $\rho = \arctan(\mathbf{x} + \delta)$ and the unrestricted approximate perturbation δ^* is as the following.

$$\delta^* = \frac{1}{2}\left(\tanh(\rho + 1)\right) - \mathbf{x} \tag{A II-8}$$

This perturbation is unrestricted and it should be tuned for feature values by measuring $\nabla f(\mathbf{x}+\delta^*)$. For feature intensities with negligible gradient values, the actual adversarial perturbation truncates to zero, and for the rest: $\delta \leftarrow \delta^*$.

Attacking victim classifiers while there is an unrestricted access to the details of the attacked models, including the training dataset, hyperparameters, architecture, and more importantly gradient information, like all the abovementioned attack algorithms, is less challenging compared to the black-box attack scenarios. Usually, in the latter scheme, the adversary runs gradient estimation via querying the classifier by training a surrogate model. In this paper, the chosen black-box attack is the natural evolution strategy (NES (Wierstra, Schaul, Peters & Schmidhuber, 2008)) which has been employed for gradient approximation in (Ilyas *et al.*, 2018). This iterative algorithm is known as partial information attack (PIA) and it encodes the l_∞ similarity metric as part of its targeted optimization problem. Finding the proper adversarial perturbation bound for PIA is to some extent challenging and requires a very high number of queries to the victim model.

Before discussing how adversarial attack and adversarially training on various deep neural network architectures have been implemented, we firstly need to provide a brief overview on the transformation of an audio signal into 2D representations. The next section will describe spectrogram generation using short time Fourier transformation (STFT), discrete wavelet transformation (DWT), and tonnetz feature extraction. We will then train our classifiers using these representations and investigate how adversarially training impacts their robustness to adversarial attacks.

4. Audio Transformation

Since audio and speech signals have high dimensionality in time domain, their 2D representations with lower dimensionalities have been widely used for training advanced classifiers originally developed for 2D computer vision applications (Esmaeilpour *et al.*, 2020c). In this work, we use STFT and DWT, both with and without tonnetz features for generating 2D representations of audio signals. This section briefly reviews the required transformations by this work.

For a discrete signal $a[n]$ distributed over time n using the Hann window function $H[\cdot]$, we can compute the complex Fourier transformation using the following equation:

$$\text{STFT}\left\{a[n]\right\}[m,\omega] = \sum_{n=-\infty}^{\infty} a[n]H[n-m]e^{-j\omega n} \tag{A II-9}$$

where m is the time scale and $m \ll n$. Additionally, ω stands for the continuous frequency coefficient. This transformation applies on shorter overlapping sub-signals with a predefined sampling rate and forms the STFT spectrogram as shown in Eq. A II-10.

$$\text{SP}_{\text{STFT}}\left\{a[n]\right\}[m,\omega] = \left|\sum_{n=-\infty}^{\infty} a[n]w[n-m]e^{-j\omega n}\right|^2 \tag{A II-10}$$

There are several variants of the STFT transformation such as mel-scale and cepstral coefficient, producing even lower dimensionality, that have been widely used for various speech processing tasks (Patel & Rao, 2010; Juvela, Bollepalli, Wang, Kameoka, Airaksinen, Yamagishi & Alku, 2018). However in this work, we use the standard STFT representation for training the front-end dense classifiers.

Generating DWT spectrogram is very similar to the Fourier transformation as they both employ continuous and differentiable basis functions. For the wavelet transformation, several functions have been studied and their effectiveness for audio signals have been investigated in (Mitra & Wang, 2008; Patidar & Pachori, 2014). The general form of this transformation for a

continuous function $a(t)$ is shown in Eq. A II-11.

$$\mathrm{DWT}\left\{a(t)\right\} = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} a(t)\psi\left(\tfrac{t-\tau}{s}\right) dt \qquad \text{(A II-11)}$$

where τ and s refer to the time variations in the transformation and the wavelet scale, respectively. Moreover, ψ stands for the basis mother functions. Common choices for this function are Haar, Mexican Hat, and complex Morlet. The latter has been extensively used in signal processing, mainly because of its nonlinear characteristics (Esmaeilpour *et al.*, 2020c) (see Eq. A II-12).

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega t} e^{-t^2/2} \qquad \text{(A II-12)}$$

The complex Morlet is continuous in its conjugate manifold. The convolution of this function with overlapping chunks of the given audio signal results in its spectral visualization as described in Eq. A II-13.

$$\mathrm{SP}_{\mathrm{DWT}}\left\{a[n]\right\} = \left|\mathrm{DWT}\left\{a[k,n]\right\}\right| \qquad \text{(A II-13)}$$

where k and n are integer numbers associated with scales of ψ.

The two aforementioned transformations represent spatiotemporal modulation features of a signal in the frequency domain, regardless of its harmonic characteristics. It has been demonstrated that using harmonic change detection function (HCDF) provides distinctive features for the audio signal (Harte, Sandler & Gasser, 2006). This function provides chromagram from the constant-Q transformation (CQT) which are also known as tonnetz features. According to (Harte *et al.*, 2006), there are four major steps in a HCDF system. Firstly, the audio signal is converted into logarithmic spectrum vectors using CQT. Then, pitch-class vectors are extracted from the tonal transformation based on the quantized chromagram. In the third step, 6-dimensional centroid vectors form a tensor from the tonal transformation. Finally, a smoothing operation postprocesses this tensor for distance calculation.

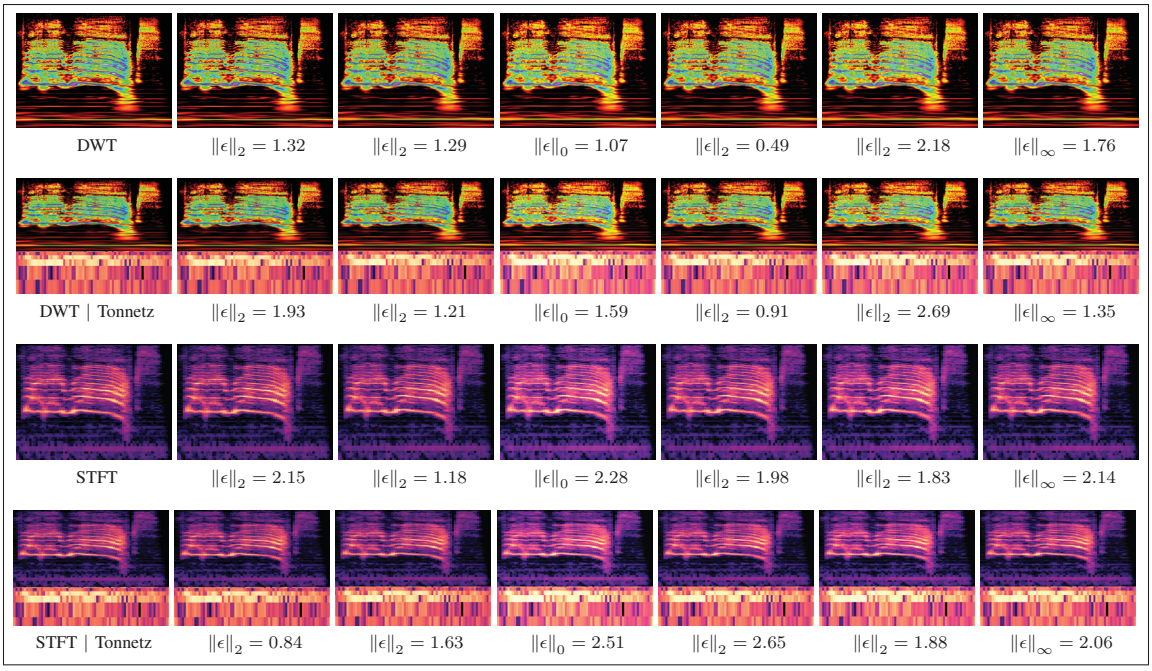

Figure 2.1 Crafted adversarial examples for the ResNet-56 using the six optimization-based attack algorithms. The first column of the figure denotes the original representations for the randomly selected sample from the class of 'children playing' in the UrbanSound8K dataset. Other columns are associated with the attack algorithms namely, BIM-a, BIM-b, JSMA, DeepFool, CWA, and PIA, respectively. Adversarial Perturbation values have been written at the bottom of each adversarial spectrogram.

We use HCDF system for generating spectrogram from audio signals in order to enhance recognition performance of the classifiers. In the next section, we provide details of this process for two benchmarking environmental sound datasets.

5. Spectrogram Production

We produce STFT representation based on the instructions provided by the open source Python library Librosa (McFee *et al.*, 2015b). We set the windows size and the hop length (n and m in Eq. A II-9) to 2048 and 512, respectively. Additionally, we initialize the number of filters to 2048 which is the standard value for the environmental sounds task (Esmaeilpour *et al.*, 2020c). Audio chunks associated with each window are padded in order to reduce the potential negative effect of losing temporal dependencies. Furthermore, the frames are overlapped using a ratio of 50%.

For generating DWT spectrograms, we use our modified version of the wavelet sound explorer (Hanov, 2008) with the complex Morlet mother function. As proposed by (Boddapati *et al.*, 2017), we set the DWT sampling frequency to 16 KHz for ESC-50 and 8 KHz for UrbanSound8K with the uniform 50% overlapping ratio. For enhancement purposes, we use the logarithmic visualization on the generated spectrograms to better characterize high frequency areas.

For the tonnetz chromagram, we use the default settings provided by Librosa with the sampling rate of 22.05 KHz. We resize the resulting chromagrams in such a way that the result will comply with the aforementioned representations. Inspired from (Su *et al.*, 2019), we append these features to the STFT and DWT spectrograms and organize them into two additional representations. In the next section, we provide more details about the training of the front-end classifiers using these four spectrogram sets.

6. Classification Models

Since an adversary runs the adversarial attack against the classifier, the choice of the victim network architecture affects the fooling rate of the model. This issue has been studied in (Esmaeilpour *et al.*, 2020) for the advanced GoogLeNet (Szegedy *et al.*, 2015) and AlexNet (Krizhevsky *et al.*, 2012) architectures trained on DWT (with linear, logarithmic, and logarithmic real visualizations), STFT, and their pooled spectrograms (appended without overlap). Since our main objective is investigating the impact of adversarially training on advanced deep learning classifiers, we additionally include ResNet-X architectures with $X \in \{18, 34, 56\}$ (He *et al.*, 2016) and VGG-16 (Simonyan & Zisserman, 2015) architectures[49].

The pretrained models of these six classifiers have been used and the input and output layers have been fine-tuned[50] as described in (Esmaeilpour *et al.*, 2020). Computational hardware used for all experiments are two NVIDIA GTX-1080-Ti with 4×11 GB memory in addition to a 64-bit Intel Core-i7-7700 (3.6 GHz) CPU with 64 GB RAM[51]. We carry out our experiments using the

[49] Input and output layers are justified to spectrograms.

[50] Without clipping weight vectors.

[51] In parallel, however without overclocking.

five-fold cross validation setup for all the spectrogram sets. As a common practice in model performance analysis, we preserve 70% of the entire samples for training and development followed by running the early stopping scenario. We report recognition accuracy of these models for the remaining 30% samples.

In the next section, we provide the detailed setup for the adversarial algorithms mentioned in section 3. We additionally discuss budget allocations required by the adversary for successfully attacking the six finely trained victim models.

7. Adversarial Attack Setup

For effectively attacking the classifiers, the adversary should tune the hyperparameters required by the attack algorithms such as the number of iteration, the perturbation limitation, the number of line search within the manifold, which we express them all as the budget allocations. For finding the optimal required budgets, we bind the fooling rates of the attack algorithms to a predefined threshold $AUC > 0.9$ associated with the area under curve of the attack success. In other words, we allocate as much budget as needed for reaching the $AUC > 0.9$ for all attacks against the victim models. This is a critical threshold for demonstrating the extreme vulnerability of neural networks against adversarial attacks.

In accordance to the above note, we use Foolbox (Rauber *et al.*, 2017), the freely available python package in support of the uniform reproducible implementations of the attack algorithms. For the BIM-a and BIM-b algorithms, we define the $\epsilon \geq 0.0015$ with the confidence of ($\geq 75\%$). In the JSMA framework, we set the number of iterations to a maximum of 1000 and the scaling factor within $[0, 250]$ (with equivalent displacement of 50). The number of iterations in the DeepFool attack is initialized to 100 with the supremum value in light of 600 and the static step of 100. For the costly CWA attack, we set the search step $c \in \{1, 3, 5, 7\}$ within the number of iteration $\{25, 100, 1\text{k}, 1.5\text{k}\}$ associated with every c. Except for the DeepFool, which is a non-targeted attack, we randomly select targeted wrong labels for the rest of the algorithms.

There are four hyperparameters required for the black-box PIA algorithm. We empirically limit the perturbation bound to $\epsilon \geq 0.001$ followed by an iterative line search to find the most approximately optimal variance in the NES gradient estimation. We initialize the number of iterations to 500 with decay rate of 0.001 and the learning rate $\eta \in [0.001, 0.6]$.

In the framework which we attack the front-end audio classifiers, we run the algorithms on the shuffled batches of 500 samples up to 50 batches of 100 samples randomly selected from the clean spectrograms in every step toward spanning the entire datasets. These attacks are performed considering the abovementioned allocated budgets once before and after adversarially training in order to measure the robustness of the models. Section 8 provides details on how adversarially training has been implemented.

8. Adversarially Training

The idea of adversarially training was firstly proposed in (Goodfellow *et al.*, 2015), where authors showed that augmenting the training dataset with the one-shot FGSM adversarial examples improves the robustness of the victim models. As commonly known, the main advantage of this simple approach is that it does not shatter nor obfuscate gradient information while runs a fast non-iterative procedure. This has made the adversarially training to be a relatively reliable defense approach. However, it may not confidently defend against stronger white-box adversarial algorithms (Tramèr *et al.*, 2017).

Many adversarial defense approaches have been introduced during the past years which have been reported to outperform FGSM-based adversarially training (Papernot, McDaniel, Wu, Jha & Swami, 2016a; Buckman, Roy, Raffel & Goodfellow, 2018; Guo, Rana, Cisse & Van Der Maaten, 2017). However, some studies have been reported that these advanced defense approaches shatter gradient vectors and they might easily break against strong adversarial attacks which do not incorporate the exact gradient information such as the backward pass differentiable approximation (Athalye *et al.*, 2018b).

Augmenting the clean training dataset with adversarial examples in the adversarially trained framework is shown in Eq. A II-14 (Goodfellow *et al.*, 2015).

$$J'(\mathbf{x}, l, \mathbf{w}) = \alpha J(\mathbf{x}, l, \mathbf{w}) + (1 - \alpha) J(\mathbf{x}', l, \mathbf{w}) \tag{A II-14}$$

where α is a subjective weight scalar definable by the adversary. Additionally, J and $\mathbf{w}$ denote the loss function and the derived weight vector of the victim model, respectively. Moreover $\mathbf{x}$ and $\mathbf{x}'$ refer to the legitimate and adversarial example associated with the genuine label l. Adversarially training using a costly attack algorithm is very time-consuming and memory prohibitive in practice. Therefore, we use the FGSM for augmenting the original spectrogram datasets with the adversarial examples according to the assumption of $J'(\mathbf{x}, l, \mathbf{w}) = J(\mathbf{x}', l, \mathbf{w})$.

In the next section, we report our achieved results for the dense neural network models about the adversarial attacks and adversarially training on four different representations, namely STFT, DWT, STFT appended with tonnetz features, and DWT appended with tonnetz chromagrams.

9. Experimental Results

We conduct our experiments on two environmental sounds datasets: UrabanSound8K (Salamon *et al.*, 2014a) and ESC-50 (Piczak, 2015b). The first dataset contains 8732 short recording arranged in 10 classes (car horn, dog bark, drilling, jackhammer, street music, siren, children playing, air conditioner, engine idling and gun shot) with the audio length of < 4 seconds. ESC-50 dataset contains 2K audio signals with an equal length of five seconds organized in 50 classes.

For enhancing both quality and quantity of these datasets, especially for ESC-50, we filter samples using the pitch-shifting operation in the temporal domain as proposed in (Esmaeilpour *et al.*, 2020c). According to their proposed 1D filtration setup, we use the scales of $\{0.75, 0.9, 1.15, 1.5\}$. This increases the size of the datasets by a factor of 4.

Table 2.1 Recognition performance (%) of the audio classifiers trained on the original spectrogram datasets (without adversarial example augmentation). Values inside of the parenthesis indicate the recognition percentage drop after adversarially training the models with the fooling rate $AUC > 0.9$. Accordingly, the maximum perturbation is achieved at $\|\epsilon\|_2 \leq 3$. Outperforming accuracies are shown in bold face.

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50	STFT	67.83, (06.89)	64.32, (10.91)	66.85, (12.13)	67.21, (14.43)	**69.77**, (09.29)	68.94, (08.32)
	DWT	70.42, (08.42)	65.39, (11.23)	67.06, (15.71)	67.55, (18.76)	**71.56**, (11.09)	71.43, (16.28)
	STFT ⌈ Tonnetz	70.11, (24.09)	64.21, (23.76)	67.62, (19.48)	66.75, (23.31)	**70.22**, (25.19)	70.18, (23.68)
	DWT ⌈ Tonnetz	68.76, (19.07)	68.31, (18.53)	68.49, (24.27)	67.15, (21.56)	**71.79**, (18.21)	68.37, (18.73)
UrbanSound8K	STFT	88.32, (10.35)	86.07, (21.43)	88.24, (14.94)	88.61, (09.19)	**88.77**, (23.06)	87.93, (14.66)
	DWT	90.10, (16.35)	87.51, (19.59)	88.07, (15.08)	88.38, (19.04)	**90.14**, (15.49)	90.11, (16.35)
	STFT ⌈ Tonnetz	88.44, (25.77)	86.81, (22.05)	88.13, (17.64)	88.38, (26.42)	89.41, (20.73)	**89.42**, (21.38)
	DWT ⌈ Tonnetz	89.32, (16.83)	87.34, (20.41)	88.76, (29.12)	89.80, (27.45)	**91.36**, (26.08)	89.97, (24.56)

Following the explanations provided in section 5 about the spectrogram production, the dimension of each resulting spectrogram is 1568×768 for both STFT and DWT (the logarithmic scale) representations on the two datasets. Moreover, the dimensions of the resulting chromagrams is 1568×540, which will be appended to the aforementioned representations. Table A-2.1 summarizes recognition accuracies of the classifiers trained on these spectrograms. Additionally, this table shows the effect of the adversarially training on the recognition performance of these models.

The classifiers in Table A-2.1 have been selected for evaluation on the test sets after running the five-fold cross-validation scenario on the randomized development portion of the training datasets. Regarding this table, different architectures of the deep neural networks show competitive performances. However, in the majority of the cases, the ResNet-56 outperforms other classifiers averaged over 10 repeated experiments on the spectrograms. The highest recognition accuracy has been achieved by the ResNet-56 architecture, trained on the appended representation of DWT and tonnetz chromagrams for both UrbanSound8K and ESC-50 datasets. The number of parameters in the ResNet-56 is 11.3% and 14.26% higher than its rival models VGG-16 and ResNet-34, respectively.

Fig. A-2.1 visually compares the adversarial examples crafted against the outperforming classifier, the ResNet-56, using the six adversarial attacks with a randomly selected audio sample and

represented with the four spectrogram approaches described earlier. Although the generated spectrograms are visually very similar to their legitimate counterparts, they all make the classifier to predict wrong labels.

Table A-2.1 also shows the drop ratio of the recognition accuracies after adversarially training the models following the procedure explained in section 8. The maximum required adversarial perturbation for complying with the fooling rate of $AUC > 0.9$ is achieved at $\|\epsilon\|_2 \leq 3$, averaged over all the attacks. In attacking the adversarially trained models, the procedure outlined in section 7 has been implemented individually for every audio classifier. According to the obtained results, adversarially training considerably reduces the performance of all models. For the ESC-50, the neural network trained on the appended representation of STFT and tonnetz features (STFT | Tonnetz) has experienced the most negative impact compared to other representations. The average drop ratio for adversarially trained models on the DWT | Tonnetz representations is slightly more than the STFT | Tonnetz counterparts for the UrbanSound8K dataset. However, for both datasets, these ratios for models trained on the DWT spectrogram are considerably higher than those trained with the STFT representations.

We measure the fooling rate of adversarially trained models after attacking them using the same six adversarial algorithms following the procedure explained in section 7 with the imposed condition of $\|\epsilon\|_2 \leq 3$ for the adversarial perturbation. This experiment uncovers the impact of adversarially training on the robustness of the audio classifiers (see Table A-2.2). We applied the aforementioned condition to make this table comparable with Table A-2.1. Regarding the results reported in Table A-2.2, adversarially training has improved the robustness of all the classifiers, particularly AlexNet.

For investigating the overall impact of the adversarially training on the robustness of audio classifiers, we attack the adversarially trained models using the same six attack algorithms without the condition of $\|\epsilon\|_2 \leq 3$. Unfortunately, we could achieve the fooling rate with $AUC > 0.9$ for all the classifiers following the attack procedure explained in section 7. However, attacking the adversarially trained models requires larger values for the adversarial perturbation

Table 2.2 Robustness comparison (average $AUC\%$) of the adversarially trained models attacked with the constraint $\|\epsilon\|_2 \leq 3$. Victim models with lower fooling rates are indicated in bold.

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50 (5-fold cross validation (avg.))	STFT	53.12	**50.97**	51.13	55.31	53.87	51.05
	DWT	55.68	**51.03**	52.56	54.18	52.26	52.23
	STFT | Tonnetz	56.18	**50.46**	53.10	55.29	54.19	52.82
	DWT | Tonnetz	55.74	**49.33**	54.87	53.77	50.42	51.37
UrbanSound8K (5-fold cross validation (avg.))	STFT	56.09	**53.24**	54.08	55.91	57.30	54.35
	DWT	58.98	**51.92**	53.59	54.40	55.86	53.66
	STFT | Tonnetz	55.80	**50.71**	52.75	51.02	54.11	52.39
	DWT | Tonnetz	58.46	**52.23**	55.13	56.81	55.38	55.26

Table 2.3 Comparison of ϵ_r for attacking the original and adversarially trained models with the constraint of $AUC > 0.9$. Higher values for ϵ_r associated with each representation are shown in bold.

Dataset	Representations	GoogLeNet	AlexNet	ResNet-18	ResNet-34	ResNet-56	VGG-16
ESC-50 (5-fold cross validation (avg.))	STFT	1.412	1.631	1.897	2.154	**2.312**	2.107
	DWT	1.562	1.509	1.741	1.982	1.976	**2.307**
	STFT | Tonnetz	1.804	1.918	2.003	**2.161**	2.095	1.674
	DWT | Tonnetz	2.014	2.336	1.788	1.903	**2.609**	2.230
UrbanSound8K (5-fold cross validation (avg.))	STFT	1.562	1.903	**2.439**	1.372	1.991	1.703
	DWT	2.154	2.287	2.764	1.644	**2.892**	1.789
	STFT | Tonnetz	2.231	2.108	1.981	2.003	1.401	**2.308**
	DWT | Tonnetz	1.606	2.199	2.405	1.604	**2.501**	1.702

($\|\epsilon\|_2$) compared to attacking the original models and consequently, increases the number of callbacks to the original spectrogram with extra batch gradient computations. This might degrade the quality of the generated spectrograms. In order to analytically compare the maximum adversarial perturbation required for the original and the adversarially trained models, we compute the average perturbation ratio as shown in Eq. A II-15:

$$\epsilon_r = \left| \frac{\epsilon_a}{\epsilon_o} \right| \tag{A II-15}$$

where ϵ_a and ϵ_o denote the average adversarial perturbation required for successfully attacking the adversarially trained and original models (both with $AUC > 0.9$), respectively. Table A-2.3 summarizes values for ϵ_r for the victim models trained on different representations. Note that an $\epsilon_r \geq 1$ indicates the positive impact of adversarially training on the robustness of the audio classifiers via increasing the computational cost of the attack by expanding the magnitude of the

required adversarial perturbation. With respect to the measured ϵ_r metric for all the front-end classifiers, the ResNet-56 architecture showed better robustness against adversarial attacks in average for 50% of the experiments. In other words, attacking this model adds additional cost for the adversary in crafting adversarial examples with the $AUC > 0.9$.

10. Conclusion

In this paper, we presented the impact of adversarially training as a gradient obfuscation-free defense approach against adversarial attacks. We trained six advanced deep learning classifiers on four different 2D representations of environmental audio signals and run five white-box and one black-box attack algorithms against these victim models. We demonstrated that adversarially training considerably reduces the recognition accuracy of the classifier but improves the robustness against six types of targeted and non-targeted adversarial examples by constraining over the maximum required adversarial perturbation to $\|\epsilon\|_2 \leq 3$. In other words, adversarially training is not a remedy for the threat of adversarial attacks, however it escalates the cost of attack for the adversary with demanding larger adversarial perturbations compared to the non-adversarially trained models.

APPENDIX III

CLASS-CONDITIONAL DEFENSE GAN AGAINST END-TO-END SPEECH ATTACKS

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in « 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) » in 2021.

Abstract

In this paper we propose a novel defense approach against end-to-end adversarial attacks developed to fool advanced speech-to-text systems such as DeepSpeech and Lingvo. Unlike conventional defense approaches, the proposed approach does not directly employ low-level transformations such as autoencoding a given input signal aiming at removing potential adversarial perturbation. Instead of that we find an optimal input vector for a class conditional generative adversarial network through minimizing the relative chordal distance adjustment between a given test input and the generator network. Then, we reconstruct the 1D signal from the synthesized spectrogram and the original phase information derived from the given input signal. Hence, this reconstruction does not add any extra noise to the signal and according to our experimental results, our defense-GAN considerably outperforms conventional defense algorithms both in terms of word error rate and sentence level recognition accuracy.

2. Introduction

The threat of adversarial attacks has been well characterized in the domains of audio and speech recognition (Schönherr, Kohls, Zeiler, Holz & Kolossa, 2018; Esmaeilpour *et al.*, 2020b). Classifiers either trained on raw signals or their corresponding 2D representations (i.e., spectrograms) are quite vulnerable against carefully crafted adversarial examples and this poses

a serious concern about safety and reliability of these models (Yakura & Sakuma, 2018). In a big picture, there are two main directions in studying adversarial attacks for speech signals: (i) generalizing strong attack algorithms developed for natural images in computer vision domain to spectrograms, taking advantage of their lower computational complexity (Koerich *et al.*, 2020; Esmaeilpour *et al.*, 2020); (ii) developing end-to-end attacks which require dealing directly with raw input signals (Carlini & Wagner, 2018; Qin *et al.*, 2019). In this paper, we focus on the latter for defense purposes since it is closely related to a black-box attack scenario in real-life applications.

Although there are different implementations for end-to-end attacks, they unanimously use variants of the logarithmic distortion metric $l_{\mathrm{dB}_{\vec{x}}}(\delta) = l_{\mathrm{dB}}(\delta) - l_{\mathrm{dB}}(\vec{x})$ (Carlini & Wagner, 2018), which measures the loudness in dB of an adversarial example $\vec{x}_{adv} = \vec{x}_{org} + \delta$ over its legitimate counterpart $\vec{x}_{org} \in \mathbb{R}^{n \times m}$, where n and m denote the length of the signal and the number of channels, respectively, and δ is the adversarial perturbation. Carlini and Wagner (Carlini & Wagner, 2018) have demonstrated the effectiveness of this measure as a constraint in their optimization formulation for attacking a speech-to-text model (C&W):

$$\min |\delta|_2^2 + \sum_i \vartheta_i . \mathcal{L}_i(\vec{x}_{org} + \delta_i, \pi_i) \quad \text{s.t.} \quad l_{\mathrm{dB}_{\vec{x}}}(\delta) < \zeta \tag{A III-1}$$

where π_i refers to a character alignment (tokens without duplication) according to the target output phrase $\mathbf{y}_i$ in such a way that $\Pr(\pi_i|\mathbf{y}_i) = \prod_j \mathbf{y}^j_{\pi^j}$. Additionally, $\mathcal{L}_i(\cdot)$ denotes the connectionist temporal classification loss (Graves *et al.*, 2006), and ϑ_i is a scaling factor. Finding an optimal value for ζ makes (A III-1) brittle since it requires searching in an exponential space for a phrase $\mathbf{p}_i$, which should reduce to π_i (after removing empty tokens). However, it has been shown that such a costly optimization formulation yields adversarial audios though sound very similar to $\vec{x}_{org}$, make the DeepSpeech system (Hannun *et al.*, 2014) generate any target phrase pre-defined by the adversary (Carlini & Wagner, 2018). Since δ is not universal, slightly perturbing $\vec{x}_{adv}$ such as playback and recording over the air might override generating such a target phrase. In response to this issue, variants of expectation over transformation (EOT) have been developed as part of the optimization formulation inspired by (Athalye *et al.*, 2018a).

Possible transformations are room impulse response, reverberation, and band-pass filters for truncating adversarial perturbation beyond human audible range (Yakura & Sakuma, 2018). However, this strong approach is more costly than (A III-1) and it fits well for short signals with a few corresponding phrases (Qin *et al.*, 2019). The improved version of EOT has been recently introduced with a minor enhancement over the aforementioned distortion metric (Qin *et al.*, 2019):

$$10\log_{10}|\rho_{\delta}|^2 - 10\log_{10}\left|\rho_{\vec{x}_{org}}\right|^2 \tag{A III-2}$$

where ρ denotes the power spectral density (PSD) function. They have also introduced a new formulation for the loss function according to the configuration of the Lingvo speech-to-text system (Shen *et al.*, 2019) :

$$\ell(\vec{x}_i, \delta_i, \mathbf{y}_i) = \mathbb{E}_{t\sim\tau}\left[\ell_{net}\left(\mathbf{y_i}^o, \mathbf{y_i}^t\right) + \alpha\ell(\vec{x}_i, \delta_i)\right] \tag{A III-3}$$

where α is a scalar and ℓ_{net} is the cross entropy loss which constrains over the normalized PSD function. Moreover, $\mathbf{y_i}^o$ and $\mathbf{y_i}^t$ denote the output and target phrases, respectively. This algorithm, which is known as robust attack, optimizes for the minimal δ_i over a set of τ transformations under varieties of room configurations. Similar minimization process has been implemented in a black-box scenario using a genetic algorithm (GA) (Taori *et al.*, 2019). Specifically, this GA-based attack (GAA) incorporates a momentum mutation approach as well as gradient estimation in order to obtain optimal candidate populations associated with a predefined target phrase.

While the fooling rate of the aforementioned adversarial attacks on DeepSpeech and Lingvo systems is almost 100%, there are few studies on defense approaches for speech-to-text systems. This might be due to the immaturity of the end-to-end attack algorithms since several playbacks of the crafted adversarial signal over the air might bypass the achieved perturbations (Qin *et al.*, 2019). Moreover, adversarial signals usually carry audible noises, even with $l_{\mathrm{dB}_{\vec{x}}}(\delta) < 0$, which makes their detection easier (Carlini & Wagner, 2018). However, reliable defense algorithms are still on demand against strong adversarial examples with less audible noises. Although there

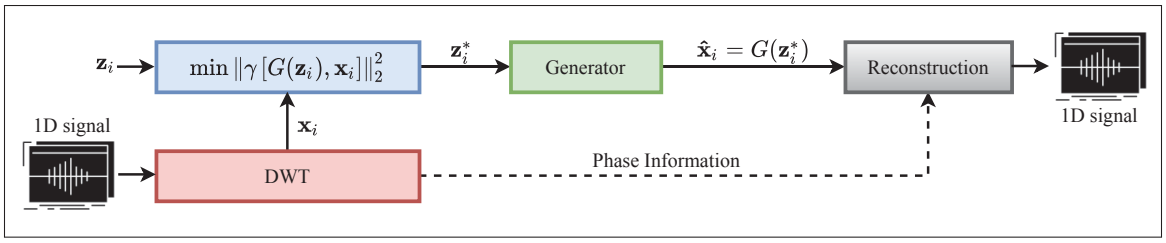

Figure 3.1 Overview of the proposed end-to-end defense-GAN approach. The 1D signal converted to a 2D-DWT spectrogram is denoted as $\mathbf{x}_i$ and the prior p_z for $\mathbf{z}_i \in \mathbb{R}^{d_z}$ is $\mathcal{N}(0, 0.4I)$. Additionally $\gamma\,[\cdot]$ is the chordal distance adjustment in the generalized Schur decomposition domain (Esmaeilpour *et al.*, 2020b) and $\hat{\mathbf{x}}_i$ represents the synthesized spectrogram from the generator. 1D signal is reconstructed using inverse DWT.

are some investigations for both proactive and reactive defense approaches (Zhang *et al.*, 2019a; Das *et al.*, 2018), they are characterized in a small scale.

In this paper, we propose a new reactive adversarial defense using a class-conditional generative adversarial network (Mirza & Osindero, 2014). We show that our proposed defense scheme can be effective for large-scale systems such as DeepSpeech and Lingvo. The rest of the paper is organized as follows. In Section 3 we provide a brief introduction to GANs focused on defense strategies for speech signals. Section 4 presents our defense approach that includes three major steps for removing potential adversarial perturbations from signals. Section 5 summarizes and discusses the experiments carried out on Mozilla common voice (MCV) and LibriSpeech datasets. The conclusion and perspective of future work are presented in the last section.

3. GAN for Adversarial Defense

In a typical GAN configuration organized as a two-player minimax optimization problem (Goodfellow *et al.*, 2014), the generator network $G(\mathbf{z}; \theta_g)$ with $\mathbf{z} \in \mathbb{R}^{d_z}$ and training parameters θ_g learns to map from the designated distribution $p_z \sim \mathcal{N}(0, I)$ to p_g as:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right] \qquad \text{(A III-4)}$$

where p_r is the real sample distribution and $D(\mathbf{x}; \theta_d)$ denotes the discriminator network with training parameters θ_d. Upon carefully training $G(\mathbf{z}; \theta_g)$, it can generate seamless samples almost without recognizable perturbations compared to $\mathbf{x}_i \sim p_r$. In fact, the generator semantically learns real sample distribution and we should expect unnoticeable differences between the generated samples and random test inputs except for adversarial examples. Based on this idea, a reactive defense approach has been introduced by Samangouei et al. (Samangouei *et al.*, 2018a), which iteratively minimizes for $\|G(\mathbf{z}) - \mathbf{x}\|_2^2$. Since the L_2 distance (or any other similarity metrics such as L_∞) between crafted adversarial examples and their corresponding legitimate samples is fairly small, they extended their optimization problem subject to finding the most optimal $\mathbf{z}_i$. Unfortunately, this adversarial filtration defense scheme shatters gradient information and it can be easily disrupted by running a backward pass differentiable approximation (BPDA) attack (Athalye *et al.*, 2018b). On the contrary, the generator network can be trained to minimize the similarity between adversarial and legitimate samples where the discriminator iteratively learns to span possible adversarial manifolds (Lee *et al.*, 2017). Training such a defense-GAN requires exploring a massive adversarial subspace since not every attack algorithm generates a universal perturbation scale (Esmaeilpour *et al.*, 2020b).

Autoencoder-based GAN (A-GAN) has also been investigated for defending speech emotion recognition models using long-short term memory networks (Latif *et al.*, 2018). This defense-GAN configuration introduces complex architecture for transforming a feature vector into another one aiming at bypassing potential adversarial perturbation. However, with the assumption of stable training without oversmoothing, this model might not necessarily enhance adversarial robustness against translation-invariant (Dong, Pang, Su & Zhu, 2019) or black-box attacks. However, these attacks are robust against low-level feature reconstruction using encoder-decoder blocks. In response to this issue and to the BPDA attack, we introduce a new defense-GAN architecture in a class-conditional framework which can be effectively used to increase the robustness of large-scale speech datasets and the state-of-the-art speech-to-text systems such as DeepSpeech and Lingvo.

4. Proposed Defense Approach: CC-DGAN

The proposed adversarial defense approach is made up of three steps, as shown in Fig. A-3.1: (i) generating signal representation; (ii) minimizing the relative chordal distance adjustment for the given input signal relative to $G(\mathbf{z}_i)$; and (iii) signal reconstruction with the preserved phase information. We explain all these steps in detail as follows.

4.1 2D Signal Representation

Due to the high dimensionality of audio and speech signals, adversarial training either on single or multi-channel waveforms is very challenging and the model often undergoes complete collapse at early iterations. Therefore, a conventional approach in speech processing is to convert a given signal into a frequency-plot representation (spectrogram). Thus, as suggested by Esmaeilpour et al. (Esmaeilpour *et al.*, 2020c), we divide the input signal into smaller chunks sampled at 16 kHz using discrete wavelet transform (DWT). Additionally, we set the frame length to 50 ms and use the complex Morlet mother function. Moreover, for enhancing the quality of the resulting spectrogram ($\mathbf{x}_i$ in Fig. A-3.1), we represent its magnitude in a logarithmic scale. It has been shown that these settings for spectrogram production outperform short-time Fourier transform both in terms of recognition accuracy and robustness against adversarial attacks (Esmaeilpour *et al.*, 2020; Esmaeilpour, Cardinal & Koerich, 2020). Since the dimensions of the generated $\mathbf{x}_i$ are not necessarily square, we bilinearly resize them to 128×128 in compliance of computing the relative chordal distance in a non-Cartesian space.

4.2 Chordal Distance Adjustment Minimization

The chordal distance (Van Loan & Golub, 1983) is a metric that measures subspace adjacency for two similar samples in the domain of generalized Schur decomposition (Esmaeilpour *et al.*, 2020b). This metric has been used for characterizing the existence of adversarial examples in subspaces different from legitimate and noisy samples (Esmaeilpour *et al.*, 2020b). The chordal distance between an adversarial example $G(\mathbf{z}_i)$ and $\mathbf{x}_i$ is:

$$\text{chord}\left(\lambda\left[G(\mathbf{z}_i)\right], \lambda\left[\mathbf{x}_i\right]\right) = \frac{\left|\lambda\left[G(\mathbf{z}_i)\right] - \lambda\left[\mathbf{x}_i\right]\right|}{\sqrt{1+\lambda\left[G(\mathbf{z}_i)\right]^2}\sqrt{1+\lambda\left[\mathbf{x}_i\right]^2}} \tag{A III-5}$$

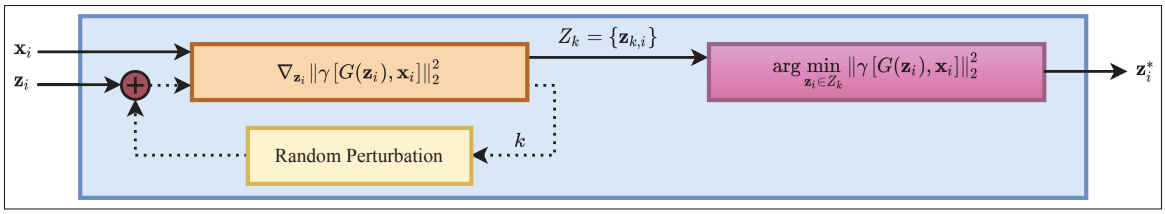

Figure 3.2 k steps minimization for the chordal distance adjustment between $G(\mathbf{z}_i)$ and $\mathbf{x}_i$. Similar to the predefined prior for $\mathbf{z}_i$, the random perturbation is also a function distributed over $\mathcal{N}(0, 0.4I)$. The inner loop is shown in dotted line.

where $\lambda[\cdot]$ denotes the vector of eigenvalues for the designated spectrograms. Achieving a valid chordal distance between two spectrograms for ensuring their subspace adjacency in the generalized Schur decomposition enquires $\|G(\mathbf{z}_i) - \mathbf{x}_i\| \simeq \xi_i$ where the threshold ξ_i must be small according to the computed mean eigenvalue. For samples which lie in the same subspace, however with dissimilar spans, a minor translation is required in Eq. A III-5 to avoid ill-conditioned cases (Van Loan & Golub, 1983). Specifically, for pencils $\vec{\mu}_i G(\mathbf{z}_i) - \mathbf{x}_i$ and $\vec{\mu}_i \in \text{diag}(\lambda\left[G(\mathbf{z}_i]\right)/\text{diag}(\lambda\left[\mathbf{x}_i\right])$, an adjustment $\gamma_i\left[\cdot\right] + \text{chord}(\cdot)$ is needed in (A III-5), especially for samples with very small L_2 distance in Euclidean space (Esmaeilpour *et al.*, 2020b).

Since the $\gamma_i\left[\cdot\right]$ adjustment is relatively large for an adversarial example $\mathbf{x}_{adv}$ (Esmaeilpour *et al.*, 2020b), minimizing over $\|\gamma\left[G(\mathbf{z}_i)\right], \gamma\left[\mathbf{x}_{adv}\right]\|_2^2$ projects $\mathbf{x}_{adv}$ onto the legitimate sample subspace distribution represented by p_g. However, we do not filter $\mathbf{x}_{adv}$, neither by conventional encoder-decoder blocks nor by low-level transformation operations. In fact, we find an optimal $\mathbf{z}_i^* \in \mathbb{R}^{d_z}$ through an iterative approach, then pass it to the generator for crafting a spectrogram very similar to the given $\mathbf{x}_{adv}$. This approach is depicted in Fig. A-3.2, where the number of iterations for obtaining the optimal $\mathbf{z}_i^*$ is denoted by k. For avoiding possible ill-conditioned pencils (Van Loan & Golub, 1983), we slightly perturb the candidate $\mathbf{z}_{k,i}$ with a random scalar and augment it with $\mathbf{z}_i$. Since $G(\mathbf{z}; \theta_g)$ is trained to support $p_g \approx p_r$, it considerably reduces the chance of generating spectrograms with adversarial perturbations. Therefore, the architectural design of both generator and discriminator has a crucial role. To this end, we propose simple yet effective class conditional architectures for reliable training.

4.2.1 Class-Conditional Defense GAN (CC-DGAN)

The proposed class-conditional defense GAN (CC-DGAN) is based on the vanilla GAN, where both the generator and the discriminator receive additional information on top of the noise vector $\mathbf{z}_i$ (i.e., $\mathbf{y}_i$) (Mirza & Osindero, 2014). Unlike the baseline model (A III-4), the CC-DGAN requires class embeddings (c-embeddings) mainly for the generator network: $\log(1 - D(G(\mathbf{z}|c = \mathbf{y})))$. This modification expands the learning space of the model at the risk of losing sample variety and mode collapse (Brock *et al.*, 2019). However, we find that c-embeddings provide a considerable boost in computing character probability distribution at every frame of the given signal compared to regular GANs.

The proposed generator receives $\mathbf{z}_i \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$ in the first layer followed by a linear block with dimension $50 + 128$ and shared $c - \text{embedding} = 50$ (Perez, Strub, de Vries, Dumoulin & Courville, 2018) including $4 \times 4 \times 16$ channels. There are two sequential residual blocks on top of the linear with $16 \rightarrow 4$ and $4 \rightarrow 1$ channels. The last hidden layer is a 128×128 non-local block with batch normalization and tanh activation function. The batch size is set 256 with orthogonal initialization (Saxe, McClelland & Ganguli, 2014). Each residual block includes two linear (128×128) and three padded convolution (3×3 with stride 1) layers followed by upsampling, batch normalization, and ReLU activation function. In our discriminator network, the first layer requires RGB spectrogram $\mathbf{x}_i \in \mathbb{R}^{128 \times 128 \times 3}$. There is only one residual block in this network which contains two sequential 3×3 convolution layers with concatenation, ReLU, skip-z (Brock *et al.*, 2019), and average pooling. On top of the residual block, there is a 64×64 non-local layer with 16 channels, ReLU, MaxPooling, and a linear logit layer ($\rightarrow 1$). Furthermore, we use both orthogonal regularization (Brock *et al.*, 2017) and initialization (Saxe *et al.*, 2014) for the entire weight vectors.

4.3 Signal Reconstruction

This is the third step of the proposed defense approach as shown in Fig. A-3.1. We reconstruct a given 1D signal with its own original phase information and the synthesized spectrogram

$\hat{\mathbf{x}}_i$. Although synthesizing phase vectors with generative models is very challenging, there are some approaches for building them. However, they add audible hissing and whining noises to the signal. Signal reconstruction with original phase vectors often provides higher signal to noise ratio and this might help to more conveniently distinguish an adversarial example from a noisy signal (Koerich *et al.*, 2020). The reconstruction operation only requires running an inverse DWT with basic settings such as type of mother function, sampling rate, and frame length. We use the same settings mentioned in Section 4.1 with additional quantization filter for normalizing the achieved vectors. For simplicity, we assume that signals are single-channel.

5. Experiments

We have evaluated the proposed defense (CC-DGAN) against three end-to-end adversarial attacks for both Mozilla's implementation of DeepSpeech (Mozilla-DeepSpeech, 2017) and Lingvo system (Shen *et al.*, 2019). These speech-to-text models are trained on Mozilla common voice (MCV) (MCV, 2019) and LibriSpeech (Panayotov *et al.*, 2015) datasets, respectively. Both these benchmarking datasets contain above 1,000 hours of voice clips with various utterances. However, as a common practice (Carlini & Wagner, 2018; Qin *et al.*, 2019; Taori *et al.*, 2019) we generate adversarial examples only for a portion of such datasets. We randomly select 11,500 and 6,000 samples from the MCV and LibriSpeech datasets for both training the CC-DGAN and running attacks, respectively. We organize these samples with their associated transcriptions into Subset-MCV and Subset-LS.

We run white-box (C&W) and black-box (GAA) adversarial attacks separately against the DeepSpeech model which uses rounds of long-short term memory blocks. We have randomly selected 1,000 samples from Subset-MCV with their original English transcriptions and we have targeted 10 different incorrect phrases (because these two attacks do not incorporate EOT) for effective attacking. Although these attacks directly optimize for achieving the minimum possible perturbation for the 1D signal, the DeepSpeech model first converts the given input into a Mel-frequency coefficient (MFC) representation. This adds more computational overhead to the attack algorithms and prohibits crafting adversarial examples for all the recordings in

the dataset. The MFC layer splits the given speech signal into 50 frames per second which means the model can output up to 50 characters per second ($\mathbf{y}_i$). Therefore, this frame length is fairly enough for short signals with quite large transcripts. We extended these two attacks for targeting silence equivalent to generating empty tokens (ϵ) for an additional 500 samples from the Subset-MCV. To this end, we updated the loss function as (Carlini & Wagner, 2018):

$$\sum_{i} \max_{t \in \{\epsilon\}} \left(f(\vec{x})^{i}_{t} - \max_{\hat{t} \notin \{\epsilon\}} f(\vec{x})^{i}_{\hat{t}}, 0 \right), \quad f : \mathcal{X}^{50} \rightarrow [0,1]^{50 \cdot |\pi|} \tag{A III-6}$$

where 50 and $\mathcal{X}$ denote the number of frames and input space, respectively. Moreover, $\hat{t}$ is the target phrase defined by the adversary in replacement of the original transcript t. Targeting ϵ token is easier than lexical characters and considerably reduces the computational cost. For the Lingvo victim model using the robust attack, we also randomly select 1,000 samples from Subset-LS with their associated transcripts targeting one incorrect phrase (because it incorporates EOT) with the same settings as mentioned in (Taori *et al.*, 2019). If the attack algorithm cannot exactly converge to a pre-defined target phrase, we replace it with another sample to keep the fooling rate at 100%.

For evaluating the proposed CC-DGAN to counteract the three adversarial attacks, we firstly train them separately on Subset-MCV and Subset-LS. In order to avoid losing sample variety and to add bias to our generative models, we exclude those nominated samples for adversarial attacks. For both generator and discriminator networks, we use Adam optimizer (Kingma & Ba, 2014) with β_1=0, β_2=0.9, and a constant learning rate $2 \cdot 10^{-4}$. We also run an exploratory search for finding the optimal number of steps required for $G(\mathbf{z}; \theta_g)$ over $D(\mathbf{x}; \theta_d)$. We eventually opted to use two steps with decay rate 0.99 on two NVIDIA GTX-1080-Ti with 4×11GB memory in addition to a 64-bit Intel Core-i7-7700 (3.6 GHz) CPU with 64GB of RAM.

As a common issue in adversarial training, the proposed CC-DGAN configuration also undergoes collapse at about 9.3k and 6.8k iterations for Subset-MCV and Subset-LS, respectively. For improving the stability of our models, we have employed spectral normalization (Miyato *et al.*, 2018) only for $G(\mathbf{z}; \theta_g)$. However, it turns out oversmoothing the generated spectrogram. For

Table 3.1 Comparison of different defense approaches against white and black-box adversarial attacks for DeepSpeech and Lingvo victim models. Better results are shown in bold face. In the robust attack, Δ is the offset scalar: $\|\delta_i\| < \zeta_i + \Delta$ (Qin *et al.*, 2019) defined by the adversary.

Model	Attack	Defense	Average k	Δ	WER (%)	SLA (%)
DeepSpeech (Subset-MCV) + 5-fold Cross Valid.	C&W	A-GAN	–	–	23.98 ± 2.14	49.17 ± 1.78
		Compression	–	–	17.41 ± 3.07	56.96 ± 2.38
		Proposed CC-DGAN	67	–	**05.37 ± 2.66**	**78.15 ± 1.08**
	GAA	A-GAN	–	–	18.54 ± 5.31	53.76 ± 3.19
		Compression	–	–	**03.81 ± 1.16**	**70.14 ± 5.72**
		Proposed CC-DGAN	54	–	03.97 ± 0.44	68.35 ± 2.51
Lingvo (Subset-LS) + 5-fold Cross Valid.	Robust Attack	A-GAN	–	300	21.23 ± 4.79	58.90 ± 2.42
		Compression	–	300	19.34 ± 3.91	54.88 ± 4.52
		Proposed CC-DGAN	59	400	**07.26 ± 3.08**	**67.36 ± 1.77**

rectifying this issue, we replaced long speech signals with shorter recordings, randomly drawn from the original datasets. The final GAN models used for further evaluations are those achieved from the checkpoints prior to potential collapse, which happens at about 10k iterations on both subsets. The k-step optimization algorithm for achieving $\mathbf{z}_i^*$ is depicted in Fig. A-3.2 and finding a minimal value for it requires generalizable generative models. Regarding our experiments, for partially unstable and somewhat oversmoothed generators, k never converges in less than 400 iterations.

For evaluating the performance of the proposed defense approach against the three aforementioned adversarial attacks, we use two metrics: (i) word error rate (WER), which is computed as $(I{+}S{+}D)/N{\times}100$ where I, S, D, and N are the total number of insertions, substitutions, deletions, and reference words, respectively (Qin *et al.*, 2019); (ii) sentence level accuracy (SLA), computed as n_c/n_{tot} where n_c is the number of samples which could achieve the correct transcript and n_c is the total number of test speech signals. Table A-3.1 summarizes the results of our experiments, where both the SLA and WER are computed for the three defense algorithms. Specifically, these two metrics measure the performance of the defenses in producing phrases which reduce to correct transcriptions for the given adversarial signals. Note that these two metrics while computed for the adversarial attacks, they measure fooling rates of the victim models in producing incorrect transcriptions as defined by the adversary. For consistent evaluation and in response to the raised concern of complete model vulnerability against end-to-end adversarial attacks (Carlini & Wagner, 2018), we set the SLA to 100% for all defense algorithms. Since for effective

evaluations we target 10 incorrect transcriptions for every speech signal under C&W and GAA attacks, the reported results are averaged over 10 different runs. Table A-3.1 shows that for the majority of the cases, the proposed CC-DGAN outperforms both simple compression and complex autoencoder-based GAN (A-GAN) in removing potential adversarial perturbations from speech signals and achieving lower WER and higher SLA. The only exception is for the GAA attack, which implements approximated gradient estimation, where simple compression achieves a slightly better performance than the proposed CC-DGAN. We noticed that for such attack, doubling k, reduces the WER in about 1.09% and increases the SLA in around 2.58% compared to k=54. For better investigating this issue, we attacked both victim models with the BPDA attack and measured the performance achieved by the proposed defense GAN. Our investigation on the same crafted adversarial examples uncovered the effectiveness of this attack on the CC-DGAN. More specifically, for reaching almost the same WER and SLA reported in Table A-3.1, k should be increased 2.37 and 3.12 times more for DeepSpeech and Lingvo systems, respectively.

6. Conclusion

In this paper, we proposed a new defense algorithm for securing advanced DeepSpeech and Lingvo systems against three end-to-end white-box and black-box adversarial attacks. The proposed CC-DGAN uses simple architectures for both the generator and discriminator with few residual blocks and a reconstructor module. This module regenerates a test input speech with the synthesized DWT spectrogram and its original phase information for seamless reconstruction. The experimental results on subsets of MCV and LibriSpeech datasets have shown that the proposed defense approach considerably outperforms other defense algorithms for the majority of the cases in terms of achieving lower WER and higher SLA. Since the performance of our defense approach is highly dependent on the generalizability of the CC-DGAN, we are inclined to improve its stability and increase its generalizability in our future studies.

APPENDIX IV

CYCLIC DEFENSE GAN AGAINST SPEECH ADVERSARIAL ATTACKS

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper Published in « IEEE Signal Processing Letters (SPL) » 2021.

Abstract

This paper proposes a new defense approach for counteracting with state-of-the-art white and black-box adversarial attack algorithms. Our approach fits in the category of implicit reactive defense algorithms since it does not directly manipulate the potentially malicious input signals. Instead, it reconstructs a similar signal with a synthesized spectrogram using a cyclic generative adversarial network. This cyclic framework helps to yield a stable generative model. Finally, we feed the reconstructed signal into the speech-to-text model for transcription. The conducted experiments on targeted and non-targeted adversarial attacks developed for attacking DeepSpeech, Kaldi, and Lingvo models demonstrate the proposed defense's effectiveness in adverse scenarios.

2. Introduction

There is a relatively increasing volume of publications on developing adversarial attacks against speech-to-text (transcription) systems in targeted and non-targeted scenarios (Carlini & Wagner, 2018; Qin *et al.*, 2019; Chen *et al.*, 2020; Schönherr *et al.*, 2020; Esmaeilpour, Cardinal & Koerich, 2021b). These attack algorithms' effectiveness has been demonstrated for the advanced DeepSpeech (Mozilla-DeepSpeech, 2017), Kaldi (Povey *et al.*, 2011), and Lingvo (Shen *et al.*, 2019) transcription systems. In general, these adversarial attacks run an optimization algorithm for $\langle \vec{x}_{\text{orig}}, \hat{\mathbf{y}}_i \rangle$ where $\vec{x}_{\text{orig}}$ stands for the original (legitimate) speech signal, and $\hat{\mathbf{y}}_i$ indicates the

associated target phrase defined by the adversary (Carlini & Wagner, 2018) (Eq. A IV-1).

$$\begin{aligned} &\min_{\delta} \|\delta\|_F + \sum_i c_i L_i(\vec{x}_{\text{adv}}, \hat{\mathbf{y}}_i) \quad \text{s.t.} \quad l_{\text{dB}}(\vec{x}_{\text{adv}}) < \epsilon, \\ &l_{\text{dB}}(\vec{x}_{\text{adv}}) = l_{\text{dB}}(\delta) - l_{\text{dB}}(\vec{x}_{\text{orig}}) \mid \vec{x}_{\text{orig}}, \vec{x}_{\text{adv}} \in \mathbb{R} \end{aligned} \tag{A IV-1}$$

where $\vec{x}_{\text{adv}} = \vec{x}_{\text{orig}} + \delta$ and δ denotes the adversarial perturbation achievable through this iterative optimization formulation. Moreover, c_i is the hyperparameter for scaling the loss function $L_i(\cdot)$ regarding the length of the ground truth phrase $\mathbf{y}_i$ ($\mathbf{y}_i \neq \hat{\mathbf{y}}_i$). Furthermore, $l_{\text{dB}}(\cdot)$ computes the relative loudness (the distortion condition) of the signal in the logarithmic dB-scale, and ϵ is the audible threshold defined by the adversary. There are several variants for Eq. A IV-1 where they often employ different loss functions, distortion conditions, and expectation over transformations (EOT).

Carlini *et al.* (Carlini & Wagner, 2018) introduced the baseline variant of the aforementioned adversarial optimization formulation (C&W attack), which incorporates the connectionist temporal classification (CTC) loss function $\mathcal{L}_i(\cdot) = L_i(\cdot)$ (Graves *et al.*, 2006). The main optimization term in this attack is:

$$\min |\delta|_2^2 + \sum_i c_i . \mathcal{L}_i(\vec{x}_{\text{orig},i} + \delta_i, \pi_i), \quad \pi_i \xrightarrow{\iota(\cdot)} \hat{\mathbf{y}}_i \tag{A IV-2}$$

where π_i refers to the tokens which eventually reduce to $\hat{\mathbf{y}}_i$ after a greedy or a beam search phrase decoding operation $\iota(\cdot)$ (Carlini & Wagner, 2018). This white-box attack is targeted, and it has been successfully characterized against the DeepSpeech transcription system. However, this algorithm is not robust against over-the-air playbacks, and it might simply bypass the optimized adversarial perturbation δ after replaying $\vec{x}_{\text{adv}}$ over a noisy environment (Carlini & Wagner, 2018; Yakura & Sakuma, 2018).

The second variant of Eq. A IV-1 was introduced by Yakura *et al.* (Yakura & Sakuma, 2018). They proposed an EOT operation to tackle the over-the-air playback issue. This operation implements

the room impulse response (RIR) filter set and extends Eq. A IV-1 to (Yakura & Sakuma, 2018):

$$\min_{\delta} \mathbb{E}_{t\in\tau,\omega} \left[\mathcal{L}(\text{mfcc}(\vec{x}_{\text{adv}}), \hat{\mathbf{y}}_i) + \alpha_t \left\|\delta\right\|\right] \quad \text{s.t.} \quad \left\|\delta\right\| < \epsilon \tag{A IV-3}$$

where α_t is a scalar for adjusting the adversarial perturbation. Furthermore, τ denotes the EOT filter set including room impulse response, and ω is the white Gaussian noise filtration operation. Both t and ω parameters contribute to capturing environmental distributions with respect to enclosed room settings. Additionally, mfcc$(\cdot)$ refers to the standard Mel-frequency cepstral coefficient transform (Davis & Mermelstein, 1980) for converting a signal into a 1D frequency-level representation. This white-box attack algorithm yields an adversarial speech signal using:

$$\vec{x}_{\text{adv}} \leftarrow \left[\vec{x}_{\text{orig}} + \Omega(\delta)\right] \circledast t + \omega \tag{A IV-4}$$

where $\circledast$ is the convolution operator, and $\Omega(\cdot)$ indicates the band-pass filtration operation for limiting the perturbation between 1 and 4 kHz. Similar to the C&W attack, the Yakura attack also uses the CTC loss function with a different distortion condition ($\left\|\delta\right\| < \epsilon$) and EOT operation. The reported results on attacking the DeepSpeech model corroborate the higher capacity of such an adversarial algorithm compared to C&W attack (Yakura & Sakuma, 2018).

Schönherr *et al.* (Schönherr *et al.*, 2020) introduced the Imperio attack, which is the third variant of Eq. A IV-1. They presented a more straightforward simulation procedure for implementing the EOT operation in a noisy environment, which essentially fits in the targeted scenario within the white-box framework. The EOT operation incorporated in the Imperio attack is adapted to transcription models using conventional learning blocks such as a hidden Markov model in the Kaldi system (Eq. A IV-5).

$$\vec{x}_{\text{adv}} = \arg\max_{\vec{x}_i} \mathbb{E}_{t\sim\tau_d} \left[P(\hat{\mathbf{y}}_i|\vec{x}_{i,t})\right] \tag{A IV-5}$$

where τ_d is a RIR filter set with adequately large dimension d, and $P(\cdot)$ denotes the logits of a simple deep neural network (DNN) used for decoding $\hat{\mathbf{y}}_i$. This attack's distortion condition is

$\|\delta\| < \epsilon_p$, where ϵ_p refers to a psychoacoustic thresholding, and the employed loss function is the cross-entropy ($\ell_{net}(\cdot)$) (Schönherr *et al.*, 2020).

The fourth variant of Eq. A IV-1 is called Metamorph, and it was proposed by Chen *et al.* (Chen *et al.*, 2020). The EOT operation incorporated in this attack is based on a novel filter set using channel impulse response (CIR) operators. CIR is fundamentally similar to the RIR but instead of only simulating room configurations, it mainly focuses on the speaker-microphone (SM) pairs' geometrical position (Eq. A IV-6).

$$\min_{\delta} \alpha_m l_{\text{dB}}(\vec{x}_{\text{adv}}) + \frac{1}{m}\mathcal{L}(\vec{x}_{\text{adv}}, \hat{\mathbf{y}}_i) \quad \text{s.t.} \quad \|\delta\| < \epsilon \tag{A IV-6}$$

where α_m makes a trade-off between the adversarial signal quality and the attack success rate, and m refers to the total number of SM pairs. This attack was primarily developed for the DeepSpeech model, and it has shown a great performance in debasing the transcription performance of such an advanced speech-to-text system.

The fifth variant of Eq. A IV-1 was developed by Qin *et al.* (Qin *et al.*, 2019). They introduced a very reliable implementation for the EOT operation, which is called the Robust Attack. Moreover, this white-box attack is targeted and uses both $\ell_{net}(\cdot)$ and a masking threshold loss function $\ell_m(\cdot)$ as follows:

$$\min_{\delta} \mathbb{E}_{t\sim\tau_c}\left[\ell_{net}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) + c_i \ell_m(\vec{x}_{\text{orig},i}, \delta_i)\right] \quad \text{s.t.} \quad \|\delta\| < \epsilon \tag{A IV-7}$$

where τ_c is the filter set defined after CIR simulations. The Robust Attack has been developed to attack the Lingvo transcription system, and the experiments have demonstrated the capability of this algorithm in crafting high-quality adversarial signals.

Developing a black-box variant for Eq. A IV-1 is very challenging since simulating RIR and CIR filter sets using the common environmental settings might not be feasible. However, there are some approximation-based attack algorithms for such an aim: the multi-objective optimization attack (MOOA) (Khare *et al.*, 2019) and the genetic algorithm attack (Taori *et al.*, 2019). These

attacks are based on achieving a surrogate model for the victim transcription system via heuristic or greedy formulation. The behavior of the loss function can be approximated by solving for an objective function with respect to the model's incorrect prediction.

This paper proposes an adversarial defense for counteracting with the adversarial attack algorithms mentioned above. In summary, this paper makes the following contributions: (i) an adversarial defense algorithm based on a cyclic generative adversarial network; (ii) novel architectures for the generator and discriminator networks; (iii) characterizing the effectiveness of our defense approach against white and black-box adversarial attacks.

3. Background: Adversarial Defense

The algorithms for defending transcription systems against adversarial attacks fit the reactive defense category to the best of our knowledge. Sallo *et al.* (Sallo, Esmaeilpour & Cardinal, 2021) proposed the only proactive defense by adversarial training for short signals. Generally, the reactive defense algorithms can be categorized into explicit and implicit subcategories.

The first subcategory includes algorithms that run low or high-level filtration operations directly on the given input speech signal to bypass (modulate) the potential adversarial perturbation. For instance, MP3 encoding and multi-rate compression (Das *et al.*, 2018) have been employed for modulating adversarial signals. These defense approaches are fundamentally inspired by Das *et al.* (Das *et al.*, 2017), and one has been demonstrated the positive impact of these low-level signal compressions on bypassing the adversarial perturbation (fading the adversarial perturbation in the entire signal). However, a similar reactive approach with a high-level signal modulation perspective has been proposed by Latif *et al.* (Latif *et al.*, 2018). This defense algorithm employs an autoencoder-based GAN (A-GAN) for reconstructing features of the given speech signal. It has been proven that both these two straightforward reactive approaches might not be able to bypass strong adversarial signals carefully crafted in enclosed environmental scenes (Esmaeilpour *et al.*, 2021a).

The second subcategory of reactive defense approaches includes algorithms that instead of low or high-level filtrations, synthesize a signal very similar to the given input speech. These approaches are inspired by Samangouei *et al.* (Samangouei *et al.*, 2018a), and they implicitly avoid potential adversarial perturbation without directly manipulating the given speech signal. The main advantage of defenses in this subcategory is their higher reliability in terms of preventing gradient vector shattering or obfuscation (see a relevant discussion in (Athalye *et al.*, 2018b)).

During the last years, generative adversarial networks (GANs), such as multi-discriminator Mel-GAN (Kumar *et al.*, 2019) and class-conditional GAN (Esmaeilpour *et al.*, 2021a), have become reliable approaches for signal synthesis. The latter generative model has been mainly developed for adversarial defense purposes and utilizes shared embeddings with multiple sequential linear and residual blocks. This approach, called class-conditional defense GAN (CC-DGAN), iteratively finds a safe input vector ($\mathbf{z}_i^*$) for the generator network via:

$$\min \|\gamma\left[G(\mathbf{z}_i), \mathbf{x}_i\right]\|_2^2, \quad \mathbf{z}_i^* \leftarrow \arg\min_{\mathbf{z}_i \in Z_k} \|\gamma\left[G(\mathbf{z}_i), \mathbf{x}_i\right]\|_2^2 \tag{A IV-8}$$

where $\gamma[\cdot]$ is an adjustment operator[52] for measuring the distance between original and adversarial signal subspaces (Esmaeilpour *et al.*, 2020b; Van Loan & Golub, 1983). $G(\cdot)$ and $\mathbf{z}_i \in \mathbb{R}^{d_z}$ denote the generator network and the random latent variable with dimension d_z, respectively. Moreover, $\mathbf{x}_i$ refers to the discrete wavelet transform (DWT[53]) spectrogram representation according to the settings mentioned in (Esmaeilpour *et al.*, 2020c). Finally, running an inverse DWT operation on $G(\mathbf{z}_i^*)$ reconstructs a high-quality signal that sounds like the input signal $\mathbf{x}_i$[54]. This defense approach has been successfully tested against the adversarial attacks mentioned in Section 2, but with relatively lower model stability and generalizability in training the generator network and demanding potentially NP-complete optimization procedure in Eq. A IV-8.

---

[52] Without clipping.

[53] With any wavelet mother function.

[54] Relative to the ground-truth.

4. Cyclic Defense GAN (CD-GAN)

The novel implicit reactive adversarial defense approach that we propose is based on a cyclic GAN, and it has three steps: converting a speech signal into a DWT spectrogram, finding a safe vector $\mathbf{z}_i^*$ for the cyclic generator network to synthesize a seamless spectrogram, and reconstructing the speech signal with an inverse DWT operation.

4.1 DWT Spectrogram

Our motivation for generating DWT spectrograms rather than using 1D speech signals or using other 2D representation is threefold: spectrograms have much lower dimensionality and fit well with DNN architectures developed for computer vision applications; DWT most likely outperforms short-time Fourier transform in terms of providing distinctive features for GANs (Esmaeilpour *et al.*, 2020c); higher stability of the GAN during training (Esmaeilpour *et al.*, 2021a).

Assuming that $a[n]$ is a discrete signal of length n, its DWT can be written as:

$$\mathrm{DWT}[\varrho, n] = 2^{\varrho/2} \sum_{\rho=0}^{n-1} a[\rho]\psi[2^{\varrho}, \rho - n] \tag{A IV-9}$$

where ϱ and ρ denote the scale and dilation hyperparameters, respectively. Moreover, ψ is the wavelet mother function, which is the complex Morlet function (Young, 2012). For obtaining the DWT spectrogram, we compute the power spectrum of this transformation as of $\mathrm{sp}_{\mathrm{DWT}} = |\mathrm{DWT}[\varrho, n]|^2$. The following subsection explains how to find safe vectors for the main generator to produce spectrograms seamless to $\mathrm{sp}_{\mathrm{DWT}}$.

4.2 Spectrogram Synthesis: Safe Vector Optimization

The overview of our proposed algorithm toward achieving a safe vector for the main generator network (G_1) is depicted in Fig. A-4.1. As shown, there are two generators (G_1, G_2) in a cyclic framework in connection with two fully dependent discriminator networks (D_1, D_2). Unlike

some conventional cyclic GANs (e.g., (Zhu *et al.*, 2017b)), we do not provide source and target inputs to the generators for mapping from one sample to another. We employ G_2 mainly as a regularizer for G_1 to tackle the stability and mode collapse issues. Concerning the superior performance of the least-square GAN (LS-GAN) configuration among generative models with symmetric divergence metrics (Hong *et al.*, 2019), we opt for this configuration for both G_1 and G_2. However, we use different settings for these networks to avoid the potential oversmoothing issue (Eq. A IV-10 (Hong *et al.*, 2019; Mao *et al.*, 2017)).

$$\min_{G_j} \frac{1}{2}\mathbb{E}_{\mathbf{z}_{j,i}\sim p_{z,j}}\left[\left(D_j(G_j(\mathbf{z}_{j,i})) - \vartheta_1\right)^2\right], \forall j \in \{1,2\} \tag{A IV-10}$$

where p_r and $p_{z,j}$ denote the real and two independent random sample distributions, respectively. Moreover, we initialize ϑ_1 to one and zero respectively for G_1 and G_2 in compliance with the standard LS-GAN configuration (Hong *et al.*, 2019). We empirically designed slightly different architectures for these generators to make a reasonable trade-off between model generalizability and stability. The main generator contains six hidden layers: a fully connected (4×4×16 channels), two stacked residuals (with 16→8 and 8→4 channels plus 512 filters), and three consecutive convolution blocks (padded with receptive fields 5×5×1 plus 256 filters) followed by batch normalization and ReLU activation function. The output layer is a transposed convolution (Mao *et al.*, 2018) with tanh, resulting in a 128×128×3 spectrogram. The second generator is more straightforward and contains three sequential 3×3×1 convolutional layers with 128 filters, skip-z through these layers (Brock *et al.*, 2019), and average pooling. The output layer of G_2 is a non-local layer with a 16→4 channel and max-pooling operation. For training the discriminator networks, we also use the standard LS-GAN configuration policy, which iteratively minimizes for (Hong *et al.*, 2019; Mao *et al.*, 2017):

$$\begin{aligned}\min_{D_j} \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_r}\left[D_j(\mathbf{x}) - 1)^2\right] + \\ \mathbb{E}_{\mathbf{z}_{j,i}\sim p_{z,j}}\left[\left(D_j(G_j(\mathbf{z}_{j,i})) - \vartheta_2\right)^2\right], \quad \forall j \in \{1,2\}\end{aligned} \tag{A IV-11}$$

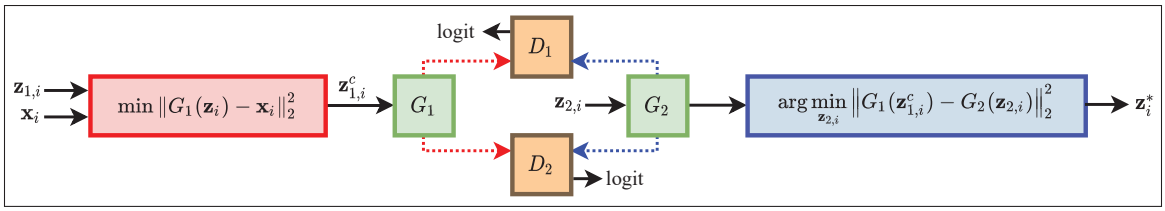

Figure 4.1 Overview of the proposed safe vector optimization procedure. G_1 (main) and G_2 are generators while D_1 and D_2 are discriminators. Herein, $\mathbf{x}_i$ stands for the input spectrogram, $\mathbf{z}_{1,i} \in p_{z,1} \sim \mathcal{N}(0, I)$, and $\mathbf{z}_{2,i} \in p_{z,2} \sim \mathcal{N}(0, 0.4I)$. Additionally, $\mathbf{z}_{1,i}^c$ and $\mathbf{z}_i^*$ indicate the candidate latent variable and the optimized safe vector, respectively.

where ϑ_2 is $\langle 0, -1 \rangle$ for D_1 and D_2, respectively. For avoiding unnecessary complications and computational overhead, we use an identical architecture for both discriminator networks. This unique architecture requires a spectrogram with a dimension 128×128×3 in the input layer on top of the five stacked hidden layers, namely two convolutions and three residuals. For the convolution blocks, we train 128 filters with receptive fields 3×3×1, followed by batch normalization and leaky ReLU activation function. On top of the residual blocks, which contain 256 filters with 4→4 and 4→1 channels, respectively, there is one non-local layer with 16 channels, max pooling, ReLU, and a linear logit layer (→1). For training our cyclic GAN, we extend the cycle-consistency loss function introduced in (Esmaeilpour *et al.*, 2020c) as:

$$\mathcal{L}_{cyclic}(\cdot) = \mathcal{L}(G_1, D_2) + \mathcal{L}(G_2, D_1) + \alpha_c \mathcal{L}(G_1, G_2) \tag{A IV-12}$$

where $0 < \alpha_c \leq 1$ is the cyclic scaling coefficient that should be empirically tuned during training. However, for simplicity and reproducibility purposes, we set this hyperparameter to 0.9.

As shown in Fig. A-4.1, we first minimize the dissimilarity between the input and the synthesized spectrograms (red rectangle) to achieve the candidate vector $\mathbf{z}_{1,i}^c$. This vector forces the main generator to yield a spectrogram seamless to $\mathbf{x}_i$. We later refine this vector by minimizing the dissimilarity between the outputs of G_1 and G_2 (blue rectangle). Upon convergence of this minimization procedure, we achieve the safe vector $\mathbf{z}_i^*$ for synthesizing the final spectrogram.

Table 4.1 Performance comparison of defense approaches against white and black-box (MOOA) adversarial attacks. Herein, reactive explicit and implicit defense algorithms are represented by RE and RI, respectively. Additionally, the maximum number of iterations before complete collapse onsets are shown and modes are computed according to (Che *et al.*, 2017). These values are averaged over 10 experiments. Outperforming values are shown in bold-face.

Transcription Model	Attack	Defense	Iteration (×10000)	Modes (×12.5)	Type	WER(%)	SLA(%)	STOI	LLR
DeepSpeech	C&W	A-GAN	01.59	0.89	RE	29.18 ± 2.1	31.63 ± 2.1	**0.84**	**0.41**
		CC-DGAN	02.67	2.55	RI	16.75 ± 3.5	60.17 ± 1.2	0.83	0.38
		CD-GAN	**02.91**	**4.52**	RI	**08.19 ± 1.3**	**71.19 ± 2.3**	0.82	0.44
	Yakura's	A-GAN	01.22	0.66	RE	20.57 ± 0.6	41.36 ± 0.4	0.83	0.35
		CC-DGAN	02.55	3.05	RI	15.97 ± 1.4	62.19 ± 1.2	**0.91**	**0.32**
		CD-GAN	**02.61**	**4.87**	RI	**11.52 ± 1.3**	**73.11 ± 2.5**	0.89	0.34
	Metamorph	A-GAN	01.04	0.71	RE	19.97 ± 1.7	56.34 ± 2.6	0.92	**0.35**
		CC-DGAN	**02.98**	**3.18**	RI	**10.26 ± 2.6**	**74.64 ± 2.8**	0.90	0.36
		CD-GAN	02.91	2.55	RI	17.42 ± 1.1	70.82 ± 2.3	**0.94**	0.41
	MOOA	A-GAN	01.27	0.54	RE	19.67 ± 3.6	50.98 ± 3.1	**0.92**	0.34
		CC-DGAN	02.89	3.76	RI	12.32 ± 1.2	62.71 ± 3.5	0.89	**0.30**
		CD-GAN	**02.94**	**4.11**	RI	**07.36 ± 2.1**	**71.11 ± 2.4**	0.91	0.35
Kaldi	Imperio	A-GAN	01.02	0.65	RE	19.58 ± 1.3	51.87 ± 2.1	0.94	0.37
		CC-DGAN	02.63	2.97	RI	12.87 ± 2.1	62.99 ± 1.3	**0.96**	**0.32**
		CD-GAN	**02.75**	**3.63**	RI	**07.49 ± 1.5**	**71.01 ± 1.9**	0.92	0.34
Lingvo	Robust Attack	A-GAN	01.02	0.56	RE	18.88 ± 1.2	58.54 ± 1.6	**0.95**	**0.30**
		CC-DGAN	02.95	2.77	RI	11.51 ± 2.3	62.58 ± 1.7	0.91	0.33
		CD-GAN	**02.96**	**3.29**	RI	**09.45 ± 1.4**	**70.96 ± 0.8**	0.94	0.34

4.3 Signal Reconstruction

The last step of our adversarial defense approach is to reconstruct the speech signal from the synthesized spectrogram using the optimized safe vector. Toward this end, we use the main generator to craft $G_1(\mathbf{z}_i^*) \mapsto \text{sp}_{\text{DWT}}^*$. This spectrogram not only is very similar to the given input spectrogram $\mathbf{x}_i$ but also does not carry the potential adversarial perturbation. For reconstructing the speech signal, we run the inverse DWT operation (Meyer, 1992) on the obtained sp_{DWT}^*.

5. Experiments

This section explains the conducted experiments on three cutting-edge transcription systems, namely DeepSpeech, Kaldi, and Lingo. These speech-to-text models are trained on MCV (MCV, 2019) and LibriSpeech (Panayotov *et al.*, 2015) datasets, which contain short ($\leq$ 6 sec) and long ($>$ 6 sec) voice recordings. We randomly select 15,000 English-speaking samples separately from these datasets, including different utterances from various ages and genders. We use

70% of these samples for training the GANs and keep the remaining portion for developing adversarial attacks, as discussed in Section 2. The main motivation for crafting adversarial signals only for a part of these datasets is following a common practice in attack development and analysis (Carlini & Wagner, 2018; Qin *et al.*, 2019; Chen *et al.*, 2020; Schönherr *et al.*, 2020; Yakura & Sakuma, 2018; Esmaeilpour *et al.*, 2021a). Furthermore, the proposed defense approach does not depend on the amount of the benchmarking samples.

We converted the training speech signal into sp_{DWT} using our modified version of the baseline wavelet explorer software (Hanov, 2008) for training our cyclic GAN. We set the DWT sampling rate to 16 kHz with a frame length of 50 ms and an overlapping ratio of 0.5. Finally, we rescale all the spectrogram to 128×128×3, matching the input layers of the generator networks. All the training and evaluation procedures were conducted on four NVIDIA GTX-1080-Ti and two Intel Core-i7-7700 (3.6 GHz, Gen. 10) with 8×11GB and 2×64GB memory, respectively.

For all the attack algorithms, we make identical assumptions for the RIR, CIR, microphone-speaker position, and room settings as discussed in Section 2. Moreover, we assign five incorrect phrases ($\hat{\mathbf{y}}_i$) to the targeted and non-targeted attacks (MOOA) randomly selected from the corresponding datasets. Finally, we compare the performance of the defense algorithms against these attacks using six objective metrics in three categories: two metrics for measuring the defense success rate; two metrics for evaluating the quality of the signals after running defenses; two metrics for assessing the generalizability and stability of the generative models. For the first category, we implemented sentence-level accuracy (SLA) and word error rate (WER) as discussed in (Qin *et al.*, 2019). According to the definitions of these metrics (Qin *et al.*, 2019), a reliable defense approach should result in higher SLA and lower WER. For the second category, we use log-likelihood ratio (LLR) (Baby & Verhulst, 2019) and short-term objective intelligibility (STOI) (Taal *et al.*, 2011), which measure the relative quality of the given signals regarding the environmental noises. These two metrics have an inverse relationship, and for a signal of higher quality, the LLR is lower. For the third category, we employ the maximum number of iterations before complete collapse onset (Brock *et al.*, 2019) and a total number of learned modes (Miyato *et al.*, 2018) for a batch size of 2×512. Table A-4.1 summarizes our

achieved results averaged over ten repeating experiments. As shown in this table, for most cases, the proposed adversarial defense approach (CD-GAN) outperforms other defense algorithms in terms of model stability (higher number of iterations before collapse onset and modes per batch) and defense success rate (lower WER and higher SLR). According to this table, there is a direct relation between model stability and defense success rate. In other words, developing more stable models most likely yields a more reliable defense approach. On the other hand, our CD-GAN often competitively fails against other defenses in terms of the quality of the reconstructed signals (lower STOI and higher LLR).

6. Conclusion

This paper introduced a novel adversarial defense algorithm against cutting-edge white and black-box as well as targeted and non-targeted speech adversarial attacks. Our defense approach is based on a cyclic GAN framework employing two generator and discriminator networks provided with the cycle-consistency CTC loss function. These networks implement layers of convolution and residual blocks for capturing local and global distributions of the training DWT spectrograms for synthesizing a reliable sample. This procedure helps to reconstruct a signal almost without adversarial perturbation. Although we have shown that our proposed CD-GAN outperforms other algorithms both in terms of model stability and defense success rate, it might not produce high-quality signals. In our future work, we will employ some regularizers on the cycle-consistency loss function based on human psychoacoustic hearing thresholding to address this issue. Moreover, we are determined to use a more comprehensive integral probability metric for training more stable GANs associated with very long and multi-speaker speech signals.

APPENDIX V

TOWARDS ROBUST SPEECH-TO-TEXT ADVERSARIAL ATTACK

Mohammad Esmaeilpour[a] , Patrick Cardinal[a] , Alessandro Lameiras Koerich[a]

[a]Le Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle (LIVIA),
Département de Génie Logiciel et des TI, École de Technologie Supérieure (ÉTS),
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper Submitted for Publication to « 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)» 2022.

Abstract

This paper introduces a novel adversarial algorithm for attacking state-of-the-art speech-to-text systems, namely DeepSpeech, Kaldi, and Lingvo. Our approach is based on developing an extension for the conventional distortion condition of the adversarial optimization formulation using the Cramér integral probability metric. Minimizing over this metric, which measures the discrepancies between original and adversarial samples' distributions, contributes to crafting signals very close to the subspace of legitimate speech recordings. This helps to yield more robust adversarial signals against playback over-the-air without employing neither costly expectation over transformation operations nor static room impulse response simulations. Our approach outperforms other targeted and non-targeted algorithms in terms of word error rate and sentence-level-accuracy with competitive performance on the crafted adversarial signals' quality. Compared to seven other strong white and black-box adversarial attacks, our proposed approach is considerably more resilient against multiple consecutive playbacks over-the-air, corroborating its higher robustness in noisy environments.

2. Introduction

During the last years and especially after the characterization of adversarial attacks for the computer vision applications (Szegedy *et al.*, 2014), several investigations have been conducted on generalizing this threat to the audio recognition and speech transcription models (Esmaeilpour

et al., 2020b; Carlini & Wagner, 2018; Qin *et al.*, 2019; Esmaeilpour *et al.*, 2020). It has been proven that adversarial signals exist for both 1D and representation (spectrogram) levels, which can seriously debase the performance of the cutting-edge speech-to-text models such as DeepSpeech (Mozilla-DeepSpeech, 2017), Kaldi(Povey *et al.*, 2011), and Lingvo (Shen *et al.*, 2019). However, developing effective adversarial signals resilient to environmental noises and room settings is challenging (Yakura & Sakuma, 2018; Szurley & Kolter, 2019). These settings include the position and characteristics of both the microphone and speaker and the room's geometry. Under various settings, simply playing the crafted adversarial signal over-the-air and recording it by another microphone most likely removes the obtained adversarial perturbation (Carlini & Wagner, 2018). For addressing this issue, several expectation over transformation (EOT) operations have been introduced (Qin *et al.*, 2019; Schönherr *et al.*, 2020; Chen *et al.*, 2020; Abdullah *et al.*, 2019). These operations often employ room filter sets (e.g., channel impulse response (Chen *et al.*, 2020)) as part of the adversarial optimization procedure to avoid bypassing the perturbation after playing over-the-air. However, developing EOT is dependent on some static room assumptions, which might negatively affect the generalizability of the filter sets (Schönherr *et al.*, 2020; Esmaeilpour *et al.*, 2021a).

In a big picture, the optimization formulation toward crafting an adversarial signal for a speech-to-text model has two parts: (i) optimization term and (ii) the distortion condition (relative constraint), as follows (Carlini & Wagner, 2018):

$$\underbrace{\min_{\delta} \|\delta\|_F + \sum_i c_i \mathcal{L}_i(\vec{x}_{\text{adv}}, \hat{\mathbf{y}}_i)}_{\text{optimization term}} \quad \text{s.t.} \quad \underbrace{l_{\text{dB}}(\vec{x}_{\text{adv}})}_{\text{distortion condition}} < \epsilon \tag{A V-1}$$

where δ is the adversarial perturbation achievable through this iterative procedure for the original input signal $\vec{x}_{\text{org}}$ to yield the adversarial signal $\vec{x}_{\text{adv}}$ ($\vec{x}_{\text{adv}} = \vec{x}_{\text{org}} + \delta$). Additionally, c_i, ϵ, and $\hat{\mathbf{y}}_i$ are the scaling coefficient, audible threshold, and the targeted incorrect phrase defined by the adversary, respectively. Furthermore, $\mathcal{L}(\cdot)$ denotes the loss function such as the connectionist temporal classification (CTC) loss (Graves *et al.*, 2006; Carlini & Wagner, 2018), the psychoacoustic loss function (Szurley & Kolter, 2019), cross entropy loss (Qin *et al.*,

2019), etc. In this typical formulation, the distortion condition is usually known as the loudness metric $l_{\mathrm{dB}}(\cdot)$ computed in the logarithmic dB-scale with respect to the human hearing range (Carlini & Wagner, 2018).

The EOT operations incorporated in the state-of-the-art adversarial attack algorithms often involve the optimization term in Eq. A V-1 (Qin *et al.*, 2019; Schönherr *et al.*, 2020; Chen *et al.*, 2020). Herein, we discuss extending the distortion condition in this equation to avoid implementing the costly EOT-based operations applied on the optimization term. This also helps to craft more robust adversarial signals. Toward this end, we review some strong adversarial attack approaches in Section 3. Then, we provide theoretical explanations on developing a relative constraint (the distortion condition) in Section 4. Finally, we analyze the achieved results from the conducted experiments on attacking speech-to-text models in Section 5. In summary, we make the following contributions in this paper:

1. developing an extension for the distortion condition of an adversarial attack formulation using the Cramér integral probability metric;
2. introducing a white-box attack framework for crafting adversarial signals more robust against over-the-air playbacks;
3. avoiding time-consuming room impulse response simulations and costly EOT operations in the adversarial optimization formulation (i.e., Eq. A V-1).

3. Background

In this section, we review the cutting-edge white and black-box adversarial attack algorithms developed against speech-to-text models. More specifically, we focus on the EOT-based attacks since they are, to some extent, capable algorithms in crafting over-the-air resilient adversarial signals (Qin *et al.*, 2019). However, we start with the baseline EOT-free C&W attack (Carlini & Wagner, 2018) developed for the DeepSpeech speech-to-text system. This algorithm is based on Eq.A V-1 and introduces a simple yet effective distortion condition for a targeted

attack scenario as the following (Carlini & Wagner, 2018).

$$l_{\mathrm{dB}}(\vec{x}_{\mathrm{adv}}) = l_{\mathrm{dB}}(\delta) - l_{\mathrm{dB}}(\vec{x}_{\mathrm{org}}) \tag{A V-2}$$

where $l_{\mathrm{dB}}(\cdot)$ can be scaled by factor of 20 to better fit the human audible range (Carlini & Wagner, 2018). The C&W attack uses the CTC loss function with an assumption of optimizing $\min_{\delta} \|\delta\|_2^2$ for the string tokens π_i (without duplication) which eventually should reduce to $\hat{\mathbf{y}}_i$ (after greedy or beam search decoding (Carlini & Wagner, 2018)). Although this distortion metric constraints the C&W algorithm to craft an adversarial signal almost seamless to the original sample $\vec{x}_{\mathrm{org}}$, it does not impose a strict condition to generate an over-the-air resilient adversarial signal. Presumably, this is due to making a reasonable trade-off between adversarial signal quality and attaining small magnitude for the adversarial perturbation δ.

The EOT operation introduced by Qin *et al.* (Qin *et al.*, 2019) uses the acoustic room simulator followed by speech reverberation filtrations for crafting resilient adversarial signals in adverse scenarios (i.e., multiple over-the-air playbacks). This algorithm is known as the Robust Attack, and it fits in the targeted adversarial category incorporating a variety of room settings for improving its performance. The optimization procedure of this attack is as follows (Qin *et al.*, 2019).

$$\min_{\delta} \mathbb{E}_{t\sim\tau}\left[\ell_{net}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) + c_i \ell_m(\vec{x}_{\mathrm{org},i}, \delta_i)\right] \quad \text{s.t.} \quad \|\delta\| < \epsilon \tag{A V-3}$$

where τ is the EOT filter set predefined (computed according to the room setting) by the adversary and $\mathbf{y}_i \neq \hat{\mathbf{y}}_i$ where the latter refers to the ground truth phrase associated with $\vec{x}_{\mathrm{org}}$. Moreover, $\ell_{net}(\cdot)$ and $\ell_m(\cdot)$ denote the cross entropy and the masking threshold loss functions, respectively. The Robust Attack has been tested on the Lingvo speech-to-text system and it has demonstrated a high capacity for crafting resilient over-the-air adversarial signals.

Yakura *et al.* (Yakura & Sakuma, 2018) introduced a similar EOT operation, which employs band-pass filtration according to the human cut-off hearing range on top of the simulated room impulse response (RIR) filter set. Moreover, this attack implements the white Gaussian noise

(WGN) filtration so as to effectively simulate the environmental noises as the following.

$$\begin{aligned} &\min_{\delta} \mathbb{E}_{t\in\tau,\omega\sim\mathcal{N}(0,\sigma^2)} \left[\mathcal{L}(\text{mfcc}(\vec{x}_{\text{adv}}), \hat{\mathbf{y}}_i) + \alpha_k \left\|\delta\right\|\right], \\ &\vec{x}_{\text{adv}} = \left[\vec{x}_{\text{org}} + \Omega(\delta)\right] \circledast t + \omega \quad \text{s.t.} \quad \left\|\delta\right\| < \epsilon \end{aligned} \tag{A V-4}$$

where ω, mfcc, and α_k denote the WGN filter drawn from the normal distribution with variance σ^2, the Mel-frequency cepstral coefficient transform (Davis & Mermelstein, 1980), and the scaling hyperparameter defined by the adversary, respectively. Additionally, $\Omega(\cdot) \in [1, 4]$ kHz refers to the band-pass filtration operation and $\circledast$ is the convolution operator. Herein, $\mathcal{L}(\cdot)$ stands for the CTC loss function, and it has been adapted to the DeepSpeech victim model. The reported experiments demonstrated that Yakura's attack outperforms the C&W in a variety of environmental scenes (Yakura & Sakuma, 2018). However, at the cost of higher computational complexity for computing the τ filter set.

One reliable approach, which implements the RIR simulation with a relatively lower computational cost is the Imperio attack (Schönherr *et al.*, 2020). This algorithm employs a deep neural network (DNN) to simulate the RIR filter set and the psychoacoustic thresholding (pst) for crafting over-the-air resilient adversarial signals (see Eq. A V-5 (Schönherr *et al.*, 2020)).

$$\vec{x}_{\text{adv}} = \underbrace{\arg\max_{\vec{x}_i} \mathbb{E}_{t\sim\tau_d} \left[P(\hat{\mathbf{y}}_i|\vec{x}_{i,t})\right]}_{\vec{x}_{\text{org}}+\kappa\left[\partial\ell_{net}(\mathbf{y},\hat{\mathbf{y}})/\partial f^*(\vec{x}_{\text{org}})\right]} \tag{A V-5}$$

where d, κ, and $f^*(\cdot)$ denote the dimension of the filter set, the learning rate and the post-activation function of the DNN model mentioned above, respectively. The EOT operation incorporated in the Imperio attack is dynamic and fits well for various room settings including meeting, lecture, and office. The distortion condition in this attack is $\delta \leq pst$ and should be tuned for every incorrect phrase $\hat{\mathbf{y}}$. Imperio has been tested on the Kaldi system. Such an attack has considerably reduced this advanced speech-to-text model's performance even after playback over-the-air.

Since the robustness of an adversarial signal over-the-air can also depend on the characteristics of both speaker and microphone, the channel impulse response (CIR) filter set is developed as part of the EOT operation in the Metamorph adversarial attack (Chen *et al.*, 2020). The general formulation of this attack is as the following.

$$\min_{\delta} \alpha_t l_{\mathrm{dB}}(\vec{x}_{\mathrm{adv}}) + \frac{1}{M}\mathcal{L}(\vec{x}_{\mathrm{org}} + \delta_i, \pi_i) \quad \text{s.t.} \quad \|\delta\| < \epsilon \tag{A V-6}$$

where α_t is the balancing coefficient between the quality of the crafted adversarial signal and the overall success rate of the attack algorithm on the victim model. Additionally, M indicates the number of microphone-speaker positions in an enclosed environment. These hyperparameters have a key role in crafting robust adversarial signals, which the adversary should precisely locate. The effectiveness of the Metamorph adversarial attack has been proven for the DeepSpeech system. However, at the cost of employing various CIR filer sets (Chen *et al.*, 2020).

Developing EOT operations for the black-box adversarial attack is extremely challenging since the adversary does not have access to the victim model and its associated settings. In response to this limitation, an over-the-line technique has been developed to surrogate the over-the-air EOT operations (Abdullah *et al.*, 2019). However, this technique requires numerous experiments to capture local and global environmental scene distributions. Regarding this concern, there are two EOT-free black-box adversarial attacks with competitive performance to the over-the-line approach in attacking the DeepSpeech system: (i) the genetic algorithm attack (GAA) (Taori *et al.*, 2019), and (ii) the multi-objective optimization attack (MOOA) (Khare *et al.*, 2019). All these algorithms are often used in targeted attack scenarios as discussed in (Esmaeilpour *et al.*, 2021a).

4. Proposed Distortion Condition & Attack

This section introduces an extension for the distortion condition of the adversarial attack formulation (Eq. A V-1) for end-to-end speech-to-text systems in targeted and non-targeted scenarios. This condition fits well for the optimization formulation of the white-box adversarial

attack scenario. Our motivation for developing such a distortion condition is threefold: improving the robustness of the adversarial speech signals after playbacks over-the-air, avoiding costly EOT operations, and keeping the quality of the crafted adversarial signal as close as possible to the ground truth input signals. Toward this end, we firstly introduce an integral probability metric (IPM) to measure discrepancies between the adversarial and original signals. Then, we build our distortion condition for adversarial attacks based on this IPM. We explain all the required details in the following subsections.

4.1 Cramér Integral Probability Metric (Cramér-IPM)

One of the standard statistical approaches in measuring the dissimilarity between two probability distributions regardless of the total number of their independent variables is using an IPM (Müller, 1997; Dodge & Commenges, 2006). Formally, an IPM is a measure for approximating the discrepancies between two (generalizable to higher orders) probability density functions $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$ as (Müller, 1997; Dedecker & Merlevède, 2007):

$$\sup_{f \in \mathcal{F}} \left[\mathbb{E}_{\vec{x}_i \sim \mathbb{P}} f(\vec{x}_i) - \mathbb{E}_{\vec{x}_i \sim \mathbb{Q}} f(\vec{x}_i) \right] \tag{A V-7}$$

where $f(\cdot)$ is called the critic function and it analytically compares the dissimilarity between $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$. Moreover, $\mathcal{F}$ denotes the possible function class for the critic and it is completely independent to both the abovementioned probability distributions (Sriperumbudur *et al.*, 2012). Mathematically, there are many choices for the function class, however we opt to Cramér ($\mathcal{F}_{Cr}$) due to its simplicity, differentiability, and generalizability (Székely, 2003; Bellemare *et al.*, 2017). The statistical definition for $\mathcal{F}_{Cr}$ in the closed form is as (Bellemare *et al.*, 2017; Rizzo & Székely, 2016; Cramér, 1928):

$$\mathcal{F}_{Cr} = \left\{ f_\vartheta : \mathcal{X} \to \mathbb{R}, \mathbb{E}_{\vec{x}_i \sim \mathbb{P}} \left(D^{(1)} f_\vartheta(\vec{x}_i) \leq 1 \right) \right\} \tag{A V-8}$$

where $D^{(1)}$ indicates the first-order derivation operator and the critic function f_ϑ is smooth with the zero boundary condition (Székely & Rizzo, 2013). Moreover, $\vec{x}_i \in \mathbb{R}^{n \times m}$ is an m-channel

signal with the length n and $\mathcal{X}$ is a compact subset in $\mathbb{R}$. According to this definition, $\mathcal{F}_{Cr}$ restricts the derivative of $f_\vartheta(\cdot)$ within a unit ball to enforce its continuity for higher degrees of ϑ (Bellemare *et al.*, 2017; Cramér, 1928).

Assuming the probability distribution functions for the original and adversarial signals are represented by $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$. Therefore, minimizing over Eq. A V-7 using the $\mathcal{F}_{Cr}$ reduces dissimilarities between random pairs of $\vec{x}_{\text{adv}}$ and $\vec{x}_{\text{org}}$. However, this minimization procedure's convergence is highly dependent on the availability of f_ϑ. One possible approach for finding this critic function could be training a neural network (mainly in the generative model frameworks (Bellemare *et al.*, 2017; Salimans, Zhang, Radford & Metaxas, 2018)), Nevertheless, it imposes unnecessary complications and computational overhead to the adversarial optimization formulation. To tackle this issue, we empirically approximate f_ϑ with the joint cumulative distribution function (CDF (Deisenroth, Faisal & Ong, 2020)) of $\mathbb{P}(\cdot)$ and $\mathbb{Q}(\cdot)$ as the following.

$$f_{\mathbb{PQ}}(\cdot) \simeq \sum_{i=1}^{n_t} \mathbb{P}(\vec{x}_{i,\text{org}}) + \mu \cdot \mathbb{Q}(\vec{x}_c), \quad \mu \sim \mathcal{U}[-1, 1] \tag{A V-9}$$

where $\vec{x}_c$ is a candidate for the adversarial signal $\vec{x}_{\text{adv}}$ achieved through optimizing for Eq. A V-1 and eventually $\vec{x}_c \xrightarrow{\mu} \vec{x}_{\text{adv}}$. Furthermore, n_t refers to the total number of original samples and μ is a uniform scaling probability prior to avoid dominating $\mathbb{P}(\vec{x}_{i,org}), \forall i$ over $\mathbb{Q}(\vec{x}_c)$. Using the critic function $f_{\mathbb{PQ}}(\cdot)$ in Eq. A V-8 provides a meaningful space for measuring discrepancies between original and adversarial distributions (see similar note in (Mroueh *et al.*, 2018)). Thus, minimizing over Eq. A V-7 maps $\vec{x}_c$ onto the original signal manifold and yields a more robust adversarial signal (We discuss this claim in Section 5).

4.2 Distortion Condition Using the Cramér-IPM

In this subsection, we introduce our distortion condition based on the Cramér-IPM with the critic function $f_{\mathbb{PQ}}(\cdot)$. In fact, we extend the relative constraint mentioned in Eq. A V-1 to:

$$\min_{\delta, f_{\mathbb{PQ}} \in \mathcal{F}_{Cr}} \left\| \mathbb{E}_{\vec{x}_i \sim \mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i) - \mathbb{E}_{\vec{x}_c \sim \mathbb{Q}} f_{\mathbb{PQ}}(\vec{x}_c) \right\| \tag{A V-10}$$

where $l_{\mathrm{dB}}(\vec{x}_c) < \epsilon$ and $\vec{x}_{\mathrm{adv}} = \arg\min \vec{x}_c$. The intuition behind exploiting this condition is finding the most possibly optimal signal $\vec{x}_c$ which not only sounds similar to $\vec{x}_{\mathrm{org}}$ according to the loudness metric $l_{\mathrm{dB}}(\cdot)$, but also lies closer to the original signal manifold. Since $\mathbb{E}_{\vec{x}_i \sim \mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i)$ incorporates the CDF of original and adversarial signals containing background and room noises, it implicitly learns the impulse responses available in the speech dataset. This also possibly makes bypassing δ very challenging after playbacks over-the-air.

From a statistical point-of-view, the proposed distortion condition forces an attack optimization formulation to craft an adversarial signal marginally close to the original signals' distribution. This is for counteracting with adversarial defense algorithms, which measure the distance between distribution manifolds to detect an adversarial signal (Esmaeilpour *et al.*, 2021a). These defense approaches are inspired by Ma *et al.* (Ma *et al.*, 2018), where it proves the subspace of adversarial signals is distinct from original and noisy samples (Esmaeilpour *et al.*, 2020b). In other words, it is possible to measure the distance between subspaces using metrics defined in orthogonal decomposition forms (e.g., chordal distance in Schur decomposition space (Esmaeilpour *et al.*, 2020b).) Based on this finding, variants of defense algorithms have been developed and they have shown a great performance against strong white, and black-box adversarial attacks (Esmaeilpour *et al.*, 2021a). Therefore, incorporating our proposed distortion condition into the attack optimization formulation (i.e., Eq. A V-1) helps to yield a more robust adversarial signal.

The general overview of our proposed attack algorithm is shown in Algorithm V-1. Regarding this pseudocode, we do not employ any EOT operations in our optimization formulation since Eq. A V-10 implicitly captures local and global distributions of the signals available in the comprehensive speech datasets.

5. Experiments

This section discusses the performance of our proposed adversarial attack algorithm, which employs the extended distortion condition using the Cramér-IPM. We implement Algorithm V-1

Algorithm-A V-1 Robust adversarial attack with distortion condition using the Cramér-IPM

Algorithm-A: Our proposed adversarial attack algorithm against speech-to-text transcription systems.

Input: $\vec{x}_{\text{org}}$, $\mathbf{y}$, $\hat{\mathbf{y}}$, ϵ

Output: $\vec{x}_{\text{adv}}$

$\vec{x}_c \leftarrow \vec{x}_{\text{org}}$; /* initializing */

initialize μ ; /* latent variable */

while $\hat{\mathbf{y}} = \mathbf{y}$ **do**

  $\delta \leftarrow \min_{\delta} \|\delta\|_F + \sum_i c_i \mathcal{L}_i(\vec{x}_c, \hat{\mathbf{y}}_i)$

  $\vec{x}_c \leftarrow \vec{x}_c + \delta$

  while $l_{dB}(\vec{x}_c) > \epsilon$ **do**

    draw a random $\mu \sim \mathcal{U}[-1, 1]$

    $\delta \leftarrow \min_{\delta, f_{\mathbb{PQ}}} \left\| \mathbb{E}_{\vec{x}_i \sim \mathbb{P}} f_{\mathbb{PQ}}(\vec{x}_i) - \mathbb{E}_{\vec{x}_c \sim \mathbb{Q}} f_{\mathbb{PQ}}(\vec{x}_c) \right\|$

    $\vec{x}_c \leftarrow \vec{x}_c + \delta$

  end while

end while

$\vec{x}_{\text{adv}} \leftarrow \vec{x}_c$; /* the adversarial signal */

to attack DeepSpeech (Mozilla's implementation), Kaldi, and Lingvo speech-to-text models without using neither RIR nor CIR filter sets.

Although the proposed algorithm resembles a targeted adversarial attack and requires defining an incorrect target phrase ($\hat{\mathbf{y}}_i$), it is generalizable to the non-targeted scenario with the assumption of choosing a random phrase for $\hat{\mathbf{y}}_i$ other than the ground-truth ($\mathbf{y}_i$). Regarding the common practice in the evaluation of adversarial attack developments that craft adversarial signals only for a portion of the given speech datasets (Carlini & Wagner, 2018; Qin *et al.*, 2019; Schönherr *et al.*, 2020; Chen *et al.*, 2020; Esmaeilpour *et al.*, 2021a), we also randomly select 1000 samples from Mozilla common voice (MCV (MCV, 2019)) and LibriSpeech (Panayotov *et al.*, 2015) to evaluate the performance of our proposed attack. These two datasets are comprehensive collections containing utterances from different genders, accents, and ages in short and long speech recordings. We equally assign ten incorrect targeted and non-targeted phrases ($\hat{\mathbf{y}}_i$) toward crafting $\vec{x}_{\text{adv}}$, for every selected signal $\vec{x}_{\text{org}}$.

Since the implementations of the benchmarking speech-to-text models are different, thus for attacking these systems we use the CTC loss function ($\mathcal{L}(\cdot)$) for DeepSpeech, and the cross-entropy loss with masking threshold ($\ell_{net}(\cdot)$, $\ell_m(\cdot)$) for the Lingvo and Kaldi systems as explained in (Qin *et al.*, 2019; Schönherr *et al.*, 2020). The rest of the settings such as defining ϵ and beam search decoding for output phrases (both $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$) follow the instructions explained in (Carlini & Wagner, 2018). We make the same assumptions in all experiments for a fair comparison to the Robust Attack, Yakura's attack, Imperio, GAA, MOOA, and Metamorph. We implement all the attack algorithms on two machines with four NVIDIA GTX-1080-Ti and two 64-bit Intel Core-i7-7700 (3.6 GHz, Gen. 10) processors with 8×11 GB and 2×64 GB memory, respectively.

We compare the adversarial attack algorithms' performance from two points of view: (i) attack success rate and (ii) adversarial signal quality. For addressing the first view, we measure the word error rate (WER) and sentence level accuracy (SLA) metrics as they have been characterized for such an aim (Qin *et al.*, 2019; Derczynski, Ritter, Clark & Bontcheva, 2013):

$$\begin{aligned} \text{WER} &= \frac{(D+I+S)}{N} \times 100 \\ \text{SLA} &= \frac{n_c}{n_{tot}} \times 100 \end{aligned} \tag{A V-11}$$

where the total number of deletions, insertions, substitutions, and reference phrases have been represented by D, I, S, and N, respectively. Moreover, n_c denotes the number of adversarial signals which they could successfully attain the same predefined phrase $\hat{\mathbf{y}}$ after passing through the speech-to-text model. Additionally, n_{tot} indicates the total number of samples.

For addressing the second view, we use three quality metrics: segmental signal to noise ratio (segSNR) (Baby & Verhulst, 2019), short-term objective intelligibility (STOI) (Taal *et al.*, 2011), and log-likelihood ratio (LLR) (Baby & Verhulst, 2019). The first two metrics compute the absolute quality of the crafted adversarial signals relative to the available ground-truth speech signals ($\vec{x}_{\text{org}}$). The main motivation behind using these two objective metrics is a realistic measurement of adversarial signal quality since not necessarily a robust adversarial attack yields

Table 5.1 Performance comparison of the adversarial algorithms for attacking speech-to-text models. Values shown for every metric are averaged over 10 experiments with different $\hat{\mathbf{y}}_i$. Types of the attacks (targeted or non-targeted) are represented by T and NT, respectively. Additionally, the EOT-based algorithms are check-marked. Herein, n_{ota} stands for the total rounds of robustness against consecutive over-the-air playbacks using static positions for the pairs of speaker and microphone. Outperforming results are shown in bold.

Transcription Model	Attack	WER (%)	SLA (%)	segSNR	STOI	LLR	Type	EOT	n_{ota}
DeepSpeech	C&W (Carlini & Wagner, 2018)	78.94 ± 2.01	30.74 ± 3.16	**21.34**	0.86	0.35	T	–	0
	Yakura's attack (Yakura & Sakuma, 2018)	80.28 ± 3.14	35.49 ± 0.28	19.57	0.82	0.38	T	✓	3
	Metamorph (Chen *et al.*, 2020)	72.48 ± 1.06	45.84 ± 4.71	17.66	0.84	0.36	T	✓	1
	GAA (Taori *et al.*, 2019)	65.80 ± 2.55	48.35 ± 3.38	17.02	0.79	0.31	T	–	1
	MOOA (Khare *et al.*, 2019)	68.06 ± 2.71	47.01 ± 1.42	18.46	0.81	0.42	T/NT	–	1
	Proposed	**88.19 ± 3.15**	**21.69 ± 3.09**	18.88	**0.88**	**0.29**	T/NT	–	**4**
Kaldi	Imperio (Schönherr *et al.*, 2020)	69.34 ± 0.47	31.49 ± 1.36	**24.71**	0.91	0.28	T	✓	2
	Proposed	**83.51 ± 1.44**	**25.86 ± 1.94**	23.16	**0.93**	**0.27**	T/NT	–	**3**
Lingvo	Robust Attack (Qin *et al.*, 2019)	84.37 ± 2.07	28.21 ± 2.31	19.44	**0.85**	**0.41**	T	✓	3
	Proposed	**89.73 ± 1.75**	**22.78 ± 2.62**	**21.58**	0.82	0.43	T/NT	–	**5**

a noise-free speech sample. In other words, the crafted $\vec{x}_{\text{adv}}$ should naturally sound like $\vec{x}_{\text{org}}$, which might carry environmental, echo, and hissing noises. Therefore, higher values for segSNR and STOI metrics interpret as the closer quality of $\vec{x}_{\text{adv}}$ to the original signals. Since these two metrics are not necessarily bounded, comparing adversarial signals' quality may not be tangible enough. We use the LLR, which is scaled between zero and one, in response to this potential concern. There is an inverse relationship between the magnitude of this metric and the quality of the signals. In other words, for adversarial signals close to their associated $\vec{x}_{\text{org}}$, the LLR is fairly low.

Table A-5.1 summarizes our achieved results. As shown in this table, our proposed attack algorithm outperforms the other algorithms in terms of WER and SLA. However, it partially fails against C&W, Imperio, and the Robust Attack in terms of quality of the crafted adversarial signals. Table A-5.1 also demonstrates that the proposed attack algorithm's robustness is higher than others after multiple consecutive playbacks over-the-air.

6. Conclusion

This paper introduced a new adversarial algorithm for effectively attacking the cutting-edge DeepSpeech, Kaldi, and Lingvo speech-to-text systems. Our proposed approach incorporates a

novel extension for the relative constraint of the adversarial optimization formulation to improve the crafted signals' robustness after multiple playbacks over-the-air. This extension minimizes over the Cramér-IPM between the probability distributions of the original and adversarial signals. This minimization operation projects a candidate adversarial signal onto the original speech recordings' subspace to counteract with potential defense approaches that measure the distance between subspaces. We experimentally demonstrated that the proposed white-box attack algorithm outperforms other advanced algorithms in terms of attack success rate according to WER and SLA metrics. Moreover, the crafted adversarial signals' average quality via our proposed attack is competitive to other algorithms using objective quality metrics of segSNR, STOI, and LLR. Our approach is EOT-free, and it has shown considerably higher robustness against consecutive playbacks over-the-air compared to other costly EOT-based adversarial algorithms. However, we could not achieve more than four playbacks averaged over the three victim models. We are determined to address this issue in our future works with developing more constraints on the critic function of the Cramér function class.

BIBLIOGRAPHY

Abdoli, S., Cardinal, P. & Koerich, A. L. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*, 136, 252–263.

Abdullah, H., Garcia, W., Peeters, C., Traynor, P., Butler, K. R. B. & Wilson, J. (2019). Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. *26th Annual Netw Distrib Syst Secur Symp*.

Addison, P. S. (2017). *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press.

Agrawal, D. M., Sailor, H. B., Soni, M. H. & Patil, H. A. (2017). Novel TEO-based Gammatone features for environmental sound classification. *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1809–1813.

Ahmad, S., Agrawal, S., Joshi, S., Taran, S., Bajaj, V., Demir, F. & Sengur, A. (2020). Environmental sound classification using optimum allocation sampling based empirical mode decomposition. *Physica A: Statistical Mechanics and its Applications*, 537, 122613.

Akbal, E. (2020). An automated environmental sound classification methods based on statistical and textural feature. *Applied Acoustics*, 167, 107413.

Akhtar, N. & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6, 14410–14430.

Akiyama, O. & Sato, J. (2019). Multitask learning and semisupervised learning with noisy data for audio tagging. *DCASE2019 Challenge*.

Alsouda, Y., Pllana, S. & Kurti, A. (2019). IoT-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random Forest. *Proceedings of the International Conference on Omni-Layer Intelligent Systems*, pp. 62–67.

Alzantot, M., Balaji, B. & Srivastava, M. (2018). Did you hear that? adversarial examples against automatic speech recognition. *arXiv Prepr arXiv:1801.00554*.

Anastasakos, T., McDonough, J., Schwartz, R. & Makhoul, J. (1996). A compact model for speaker-adaptive training. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 2, 1137–1140.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J. & Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*.

Arjovsky, M. & Bottou, L. (2017). Towards Principled Methods for Training Generative Adversarial Networks. *5th Intl Conf Learn Repres*.

Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Arthur, D. & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *18th annual ACM-SIAM Symposium on Discrete algorithms*, pp. 1027–1035.

Athalye, A., Engstrom, L., Ilyas, A. & Kwok, K. (2018a). Synthesizing robust adversarial examples. *Intl Conf Mach Learn*, pp. 284–293.

Athalye, A. & Carlini, N. (2018). On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*.

Athalye, A., Carlini, N. & Wagner, D. (2018b). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning*, pp. 274–283.

Aytar, Y., Vondrick, C. & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, pp. 892–900.

Baby, D. & Verhulst, S. (2019). Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. 106–110.

Bahdanau, D., Cho, K. & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. 4945–4949.

Bao, R., Liang, S. & Wang, Q. (2018). Featurized bidirectional gan: Adversarial defense via adversarially learned semantic inference. *arXiv preprint arXiv:1805.07862*.

Bay, H., Tuytelaars, T. & Van Gool, L. (2006). Surf: Speeded up robust features. *European conference on computer vision*, pp. 404–417.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S. & Munos, R. (2017). The Cramer Distance as a Solution to Biased Wasserstein Gradients. *CoRR*, abs/1705.10743.

Benesty, J., Chen, J. & Habets, E. A. (2011). *Speech enhancement in the STFT domain*. Springer Science & Business Media.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G. & Roli, F. (2013). Evasion attacks against machine learning at test time. *Joint Europ Conf Mach Learn Knowl Discov Datab*, pp. 387–402.

Boddapati, V., Petef, A., Rasmusson, J. & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Comp Sci*, 112, 2048–2056.

Bodini, M. (2019). Sound classification and localization in service robots with attention mechanisms. *Computer-Aided Developments: Electronics and Communication: Proceeding of the First Annual Conference on Computer-Aided Developments in Electronics and Communication (CADEC-2019), Vellore Institute of Technology, Amaravati, India, 2-3 March 2019*, pp. 69.

Brazdil, P. B. & Soares, C. (2000). A Comparison of Ranking Methods for Classification Algorithm Selection. *Machine Learning: ECML 2000*, pp. 63–75.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). Classification and Regression Trees (Wadsworth, Belmont, California). *Breiman, L. and Spector, P.(1992) Submodel selection and evaluation in regression. The X-random case. International Statistical Review*, 60, 291–319.

Brezis, H. (2010). *Functional analysis, Sobolev spaces and partial differential equations*. Springer, New York, NY.

Brock, A., Lim, T., Ritchie, J. M. & Weston, N. (2017). CNeural photo editing with introspective adversarial networks. *Intl Conf Mach Learn*.

Brock, A., Donahue, J. & Simonyan, K. (2019). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *Intl Conf Learn Repres*.

Brust, C.-A., Sickert, S., Simon, M., Rodner, E. & Denzler, J. (2015). Convolutional patch networks with spatial prior for road detection and urban scene understanding. *arXiv preprint arXiv:1502.06344*.

Buckman, J., Roy, A., Raffel, C. & Goodfellow, I. (2018). Thermometer encoding: One hot way to resist adversarial examples. *International Conference on Learning Representations*.

Cai, R., Lu, L., Hanjalic, A., Zhang, H.-J. & Cai, L.-H. (2006). A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on audio, speech, and language processing*, 14(3), 1026–1039.

Cances, L. & Pellegrini, T. (2021). Comparison of Deep Co-Training and Mean-Teacher approaches for semi-supervised audio tagging. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 361–365.

Carlini, N. & Wagner, D. (2017a). Adversarial examples are not easily detected: Bypassing ten detection methods. *Proc. of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14.

Carlini, N. & Wagner, D. (2018). Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. *IEEE Secur Privacy Works*, pp. 1-7.

Carlini, N. & Wagner, D. (2017b). Towards evaluating the robustness of neural networks. *2017 ieee symposium on security and privacy (sp)*, pp. 39–57.

Chan, W., Jaitly, N., Le, Q. & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4960–4964.

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q. & Norouzi, M. (2021). SpeechStew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*.

Chandrakala, S., Venkatraman, M., Shreyas, N. & Jayalakshmi, S. (2021). Multi-view representation for sound event recognition. *Signal, Image and Video Processing*, 1–9.

Chaudhuri, S. & Raj, B. (2013). Unsupervised hierarchical structure induction for deeper semantic analysis of audio. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 833–837.

Che, T., Li, Y., Jacob, A. P., Bengio, Y. & Li, W. (2017). Mode Regularized Generative Adversarial Networks. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Chen, T., Shangguan, L., Li, Z. & Jamieson, K. (2020). Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. *Netw Distrib Syst Secur Symp*.

Chen, Y., Guo, Q., Liang, X., Wang, J. & Qian, Y. (2019). Environmental sound classification with dilated convolutions. *Applied Acoustics*, 148, 123–132.

Cheng, S., Dong, Y., Pang, T., Su, H. & Zhu, J. (2019). Improving black-box adversarial attacks with a transfer-based prior. *arXiv preprint arXiv:1906.06919*.

Childers, D. G., Skinner, D. P. & Kemerait, R. C. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10), 1428–1443.

Chiu, L. K., Gestner, B. & Anderson, D. V. (2011). Design of analog audio classifiers with AdaBoost-based feature selection. *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pp. 2469–2472.

Chorowski, J., Weiss, R. J., Bengio, S. & van den Oord, A. (2019). Unsupervised speech representation learning using Wavenet autoencoders. *IEEE/ACM Trans Audio, Speech, Language Process*, 27(12), 2041–2053.

Chu, S., Narayanan, S. & Kuo, C.-C. J. (2009a). Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6), 1142–1158.

Chu, S., Narayanan, S. & Kuo, C.-C. J. (2009b). A semi-supervised learning approach to online audio background detection. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1629–1632.

Chui, C. K. (1993). *Wavelets: a tutorial in theory and applications*. Academic Press Professional, Inc.

Ciaburro, G. & Iannace, G. (2020). Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms. *Informatics*, 7(3), 23.

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M. & Schmidhuber, J. (2011). High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*.

Coates, A. & Ng, A. Y. (2012). Learning feature representations with k-means. In *Neural networks: Tricks of the trade* (pp. 561–580). Springer.

Cohen, G., Sapiro, G. & Giryes, R. (2020). Detecting adversarial samples using influence functions and nearest neighbors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14462.

Constantinou, C. C., Michaelides, E., Alexopoulos, I., Pieri, T., Neophytou, S., Kyriakides, I., Abdi, E., Reodica, J. & Hayes, D. R. (2021). Modeling the Operating Characteristics of IoT for

Underwater Sound Classification. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 1016–1022.

Costa, Y. M. G., Oliveira, L. E. S., Koerich, A. L., Gouyon, F. & Martins, J. G. (2012). Music genre classification using LBP textural features. *Sign Proc*, 92(11), 2723–2737.

Cotton, C. V. & Ellis, D. P. (2011). Spectral vs. spectro-temporal features for acoustic event detection. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 69–72.

Cowling, M. & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Patt Recog Lett*, 24(15), 2895–2907.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1), 13–74.

Cristani, M., Bicego, M. & Murino, V. (2004). On-line adaptive background modelling for audio surveillance. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2, 399–402.

da Silva, B., W Happi, A., Braeken, A. & Touhafi, A. (2019). Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems. *Applied Sciences*, 9(18), 3885.

Dai, W., Dai, C., Qu, S., Li, J. & Das, S. (2017). Very deep convolutional neural networks for raw waveforms. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 421–425.

Das, N., Shanbhogue, M., Chen, S.-T., Chen, L., Kounavis, M. E. & Chau, D. H. (2018). ADAGIO: Interactive Experimentation with Adversarial Attack and Defense for Audio. *arXiv preprint arXiv:1805.11852*.

Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M. E. & Chau, D. H. (2017). Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*.

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International journal for advance research in engineering and technology*, 1(6), 1–4.

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust, Speech, Signal Process*, 28(4), 357–366.

Dedecker, J. & Merlevède, F. (2007). The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in L_p. *ESAIM: Probability and Statistics*, 11, 102–114.

Deisenroth, M. P., Faisal, A. A. & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Demir, F., Turkoglu, M., Aslan, M. & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. *Applied Acoustics*, 170, 107520.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.

Deng, L. & O'Shaughnessy, D. (2018). *Speech processing: a dynamic and optimization-oriented approach*. CRC Press.

Deng, L., Hinton, G. & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8599–8603.

Derczynski, L., Ritter, A., Clark, S. & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. *Intl Conf Rec Adv Nat Lang Process*, pp. 198–206.

Dhillon, I. S. & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2), 143–175.

Dieleman, S. & Schrauwen, B. (2013). Multiscale approaches to music audio feature learning. *14th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 116–121.

Dixit, M., Kwitt, R., Niethammer, M. & Vasconcelos, N. (2017). Aga: Attribute-guided augmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7455–7463.

Dodge, Y. & Commenges, D. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Dong, Y., Pang, T., Su, H. & Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. *IEEE Conf Comput Vision Patt Recog*, pp. 4312–4321.

Du, T., Ji, S., Li, J., Gu, Q., Wang, T. & Beyah, R. (2019). SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems. *arXiv Prepr arXiv:1901.07846*.

Dziugaite, G. K., Roy, D. M. & Ghahramani, Z. (2015). Training generative Neural networks via Maximum Mean Discrepancy optimization. *31st Conf Uncert Artif Intell*, pp. 258–267.

Ellis, D. P. (2007). Classifying Music Audio with Timbral and Chroma Features. *International Society for Music Information Retrieval Conference (ISMIR)*, 7, 339–340.

Ellis, D. P. & Lee, K. (2004). Minimal-impact audio-based personal archives. *1st ACM workshop on Continuous archival and retrieval of personal experiences*, pp. 39–47.

Endo, Y. & Miyamoto, S. (2015). Spherical k-means++ clustering. *Modeling Decisions for Artificial Intelligence*, pp. 103–114.

Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G. & Huopaniemi, J. (2006). Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), 321–329.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2020). From Sound Representation to Model Robustness. *CoRR*, abs/2007.13703. Consulted at https://arxiv.org/abs/2007.13703.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2020). A Robust Approach for Securing Audio Classification Against Adversarial Attacks. *IEEE Trans Inf Forensics Security*, 15, 2147-2159.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2021a). Class-Conditional Defense GAN Against End-to-End Speech Attacks. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. to appear.

Esmaeilpour, M., Mansouri, A. & Mahmoudi-Aznaveh, A. (2013). A new SVD-based image quality assessment. *8th Iranian Conf Mach Vis Image Proc*, pp. 370–374.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2020a). Unsupervised feature learning for environmental sound classification using Weighted Cycle-Consistent Generative Adversarial Network. *Applied Soft Computing*, 86, 105912.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2020b). Detection of adversarial attacks and characterization of adversarial subspace. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3097–3101.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2020c). Unsupervised feature learning for environmental sound classification using weighted cycle-consistent generative adversarial network. *Applied Soft Computing*, 86, 105912.

Esmaeilpour, M., Sallo, R. A., St-Georges, O., Cardinal, P. & Koerich, A. L. (2020d). Conditioning Trick for Training Stable GANs. *arXiv preprint arXiv:2010.05844*.

Esmaeilpour, M., Cardinal, P. & Koerich, A. L. (2021b). Towards Robust Speech-to-Text Adversarial Attack. *arXiv preprint arXiv:2103.08095*.

Feinman, R., Curtin, R. R., Shintre, S. & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*.

Feng, Y., Wang, D. & Liu, Q. (2017). Learning to Draw Samples with Amortized Stein Variational Gradient Descent. *33rd Conf Uncert Artif Intell*.

Foggia, P., Saggese, A., Strisciuglio, N. & Vento, M. (2014). Cascade classifiers trained on gammatonegrams for reliably detecting audio events. *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 50–55.

Franzini, M., Lee, K.-F. & Waibel, A. (1990). Connectionist Viterbi training: a new hybrid method for continuous speech recognition. *International Conference on Acoustics, Speech, and Signal Processing*, pp. 425–428.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2), 75–98.

Gales, M. & Young, S. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1, 233–236.

Ganchev, T., Fakotakis, N. & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. *Proceedings of the SPECOM*, 1, 191–194.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1), 2096–2030.

Gerek, Ö. N. & Ece, D. G. (2008). Compression of power quality event data using 2D representation. *Electric Power Systems Research*, 78(6), 1047–1052.

Gil, Y., Chai, Y., Gorodissky, O. & Berant, J. (2019). White-to-black: Efficient distillation of black-box adversarial attacks. *arXiv preprint arXiv:1904.02405*.

Godino-Llorente, J. I., Gomez-Vilda, P. & Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE transactions on biomedical engineering*, 53(10), 1943–1953.

Gold, B., Morgan, N. & Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

Gonzalez, R. C. (2016). Digital image processing. Prentice hall.

Gonzalez, R. C. & Woods, R. E. (2002). Digital image processing second edition. *Beijing: Publishing House of Electronics Industry*, 455.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, pp. 2672–2680.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016). *Deep learning*. MIT Press Cambr.

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent Neural networks. *23rd Intl Conf Mach Learn*, pp. 369–376.

Graves, A., Mohamed, A.-r. & Hinton, G. (2013). Speech recognition with deep recurrent Neural networks. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. 6645–6649.

Griffin, D. & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans Acoust, Speech, Signal Process*, 32(2), 236–243.

Grosse, K., Manoharan, P., Papernot, N., Backes, M. & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, pp. 5767–5777.

Guo, C., Rana, M., Cisse, M. & Van Der Maaten, L. (2017). Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*.

Guzhov, A., Raue, F., Hees, J. & Dengel, A. (2021). Esresnet: Environmental sound classification based on visual domain models. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4933–4940.

Han, W., Chan, C.-F., Choy, C.-S. & Pun, K.-P. (2006). An efficient MFCC extraction method in speech recognition. *2006 IEEE international symposium on circuits and systems*, pp. 4–pp.

Han, W., Coutinho, E., Ruan, H., Li, H., Schuller, B., Yu, X. & Zhu, X. (2016). Semi-supervised active learning for sound classification in hybrid learning environments. *PloS one*, 11(9),

e0162075.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hanov, S. (2008). Wavelet sound explorer software.

Harte, C., Sandler, M. & Gasser, M. (2006). Detecting harmonic change in musical audio. *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pp. 21–26.

Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J. & Hansen, L. (2016). Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. *Artificial Intelligence and Statistics*, pp. 342–350.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Heittola, T., Mesaros, A., Eronen, A. & Virtanen, T. (2013). Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1), 1.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. et al. (2017). CNN architectures for large-scale audio classification. *IEEE Intl Conf Acous Speech Sign Proc*, pp. 131–135.

Hoang, Q., Nguyen, T. D., Le, T. & Phung, D. (2018). MGAN: Training Generative Adversarial Nets with Multiple Generators. *International Conference on Learning Representations*.

Hogg, R. V. & Ledolter, J. (1987). *Engineering statistics*. Macmillan Pub Co.

Homsi, M. N., Medina, N., Hernandez, M., Quintero, N., Perpiñan, G., Quintana, A. & Warrick, P. (2016). Automatic heart sound recording classification using a nested set of ensemble algorithms. *2016 Computing in Cardiology Conference (CinC)*, pp. 817–820.

Hong, Y., Hwang, U., Yoo, J. & Yoon, S. (2019). How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)*, 52(1), 1–43.

Hönig, F., Stemmer, G., Hacker, C. & Brugnara, F. (2005). Revising perceptual linear prediction (PLP). *Ninth European Conference on Speech Communication and Technology*.

Hu, H., Xu, M.-X. & Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 4, IV–413.

Hu, S., Shang, X., Qin, Z., Li, M., Wang, Q. & Wang, C. (2019a). Adversarial examples for automatic speech recognition: attacks and countermeasures. *IEEE Communications Magazine*, 57(10), 120–126.

Hu, S., Yu, T., Guo, C., Chao, W. & Weinberger, K. Q. (2019b). A New Defense Against Adversarial Images: Turning a Weakness into a Strength. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1633–1644.

Hu, Y. & Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *Journal Acoust Soc America*, 122(3), 1777–1786.

Huang, C.-y., Lin, Y. Y., Lee, H.-y. & Lee, L.-s. (2021). Defending your voice: Adversarial attack on voice conversion. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 552–559.

Huang, J. J. & Leanos, J. J. A. (2018). Aclnet: efficient end-to-end audio classification cnn. *arXiv preprint arXiv:1811.06669*.

Huang, Z., Xu, W. & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ilyas, A., Engstrom, L., Athalye, A. & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. *International Conference on Machine Learning*, pp. 2137–2146.

Ireland, D., Knuepffer, C. & McBride, S. J. (2015). Adaptive multi-rate compression effects on vowel analysis. *Frontiers in bioengineering and biotechnology*, 3, 118.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

Jang, Y., Zhao, T., Hong, S. & Lee, H. (2019). Adversarial defense via learning to generate diverse attacks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2740–2749.

Jensen, A. & la Cour-Harbo, A. (2001). *Ripples in mathematics: the discrete wavelet transform*. Springer Science & Business Media.

Jiang, L., Ma, X., Chen, S., Bailey, J. & Jiang, Y.-G. (2019). Black-box adversarial attacks on video recognition models. *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 864–872.

Juang, B.-H., Rabiner, L. & Wilpon, J. (1987). On the use of bandpass liftering in speech recognition. *IEEE Trans Acous Speech Sign Proc*, 35(7), 947–954.

Juvela, L., Bollepalli, B., Wang, X., Kameoka, H., Airaksinen, M., Yamagishi, J. & Alku, P. (2018). Speech waveform synthesis from MFCC sequences with generative adversarial networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5679–5683.

Kaneko, T., Takaki, S., Kameoka, H. & Yamagishi, J. (2017). Generative Adversarial Network-Based Postfilter for STFT Spectrograms. *INTERSPEECH*, pp. 3389–3393.

Kaur, R., Sandhu, R. S., Gera, A., Kaur, T. & Gera, P. (2020). Intelligent voice bots for digital banking. In *Smart Systems and IoT: Innovations in Computing* (pp. 401–408). Springer.

Keane, S. (2010). Banking on voice for large scale remote authentication. *Biometric Technology Today*, 2010(8), 8–10.

Khare, S., Aralikatte, R. & Mani, S. (2019). Adversarial Black-Box Attacks on Automatic Speech Recognition Systems Using Multi-Objective Evolutionary Optimization. *20th Annual Conf Intl Speech Comm Assoc*, pp. 3208–3212. doi: 10.21437/Interspeech.2019-2420.

Kim, S., Sundaram, S., Georgiou, P. & Narayanan, S. (2009). Audio scene understanding using topic models. *Neural Information Processing System (NIPS) Workshop (Applications for Topic Models: Text and Beyond)*.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocabiyikoglu, A. C., Besacier, L. & Kraif, O. (2018). Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. *arXiv preprint arXiv:1802.03142*.

Koerich, K. M., Esmailpour, M., Abdoli, S., Jr., A. S. B. & Koerich, A. L. (2019). Cross-Representation Transferability of Adversarial Attacks: From Spectrograms to Audio Waveforms. *arXiv preprint arXiv:1910.10106*.

Koerich, K. M., Esmailpour, M., Abdoli, S., Britto, A. d. S. & Koerich, A. L. (2020). Cross-representation transferability of adversarial attacks: From spectrograms to audio waveforms. *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097–1105.

Kumar, A., Khadkevich, M. & Fügen, C. (2018). Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 326–330.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y. & Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *Adv Neural Inf Process Syst*, pp. 14910–14921.

Kurakin, A., Goodfellow, I. & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv Prepr arXiv:1607.02533*.

Latif, S., Rana, R. & Qadir, J. (2018). Adversarial Machine Learning and speech emotion recognition: Utilizing generative adversarial networks for robustness. *arXiv preprint arXiv:1811.11402*.

Lee, H., Kang, W. H., Cheon, S. J., Kim, H. & Kim, N. S. (2021). Gated Recurrent Context: Softmax-Free Attention for Online Encoder-Decoder Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 710–719.

Lee, H., Han, S. & Lee, J. (2017). Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech & language*, 9(2), 171–185.

Li, C., Chang, W., Cheng, Y., Yang, Y. & Póczos, B. (2017). MMD GAN: Towards Deeper Understanding of Moment Matching Network. *Adv Neural Inf Process Syst*, pp. 2203–2213.

Li, J., Zhang, X., Jia, C., Xu, J., Zhang, L., Wang, Y., Ma, S. & Gao, W. (2020a). Universal adversarial perturbations generative network for speaker recognition. *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6.

Li, Y. & Gal, Y. (2017). Dropout inference in bayesian neural networks with alpha-divergences. *arXiv Prepr arXiv:1703.02914*.

Li, Z., Wu, Y., Liu, J., Chen, Y. & Yuan, B. (2020b). AdvPulse: Universal, Synchronization-free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1121–1134.

Lin, L., Wang, X., Liu, H. & Qian, Y. (2020). Guided learning for weakly-labeled semi-supervised sound event detection. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630.

Ling-li, L. (2011). Environmental Sound Classification Based on HMM and SVM. *Computer Era*, 11.

Liu, C., Feng, L., Liu, G., Wang, H. & Liu, S. (2019). Bottom-up Broadcast Neural Network For Music Genre Classification. *arXiv Prepr arXiv:1901.08928*.

Liu, Y., Chen, X., Liu, C. & Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv Prepr arXiv:1611.02770*.

Logan, B. et al. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *International Society for Music Information Retrieval Conference (ISMIR)*, 270, 1–11.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *7th IEEE international conference on computer vision*, 2, 1150–1157.

Lu, J., Ma, R., Liu, G. & Qin, Z. (2021). Deep Convolutional Neural Network with Transfer Learning for Environmental Sound Classification. *2021 International Conference on Computer, Control and Robotics (ICCCR)*, pp. 242–245.

Lu, L., Zhang, H.-J. & Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on speech and audio processing*, 10(7), 504–516.

Luo, H., Zhang, S., Lei, M. & Xie, L. (2021). Simplified self-attention for transformer-based end-to-end speech recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 75–81.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S. N. R., Schoenebeck, G., Song, D., Houle, M. E. & Bailey, J. (2018). Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Mallat, S. (2008). *A wavelet tour of signal processing: the sparse way*. Academic Press.

Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10), 1331–1398.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. & Smolley, S. P. (2017). Least squares generative adversarial networks. *International Conference on Computer Vision (ICCV)*, pp. 2813–2821.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. & Smolley, S. P. (2018). On the effectiveness of least squares generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12), 2947–2960.

Marchegiani, L. & Posner, I. (2017). Leveraging the urban soundscape: Auditory perception for smart vehicles. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6547–6554.

Masure, L., Dumas, C. & Prouff, E. (2019). Gradient visualization for general characterization in profiling attacks. *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pp. 145–167.

Masuyama, Y., Yatabe, K., Koizumi, Y., Oikawa, Y. & Harada, N. (2019). Deep griffin–lim iteration. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. 61–65.

Mathur, A., Isopoussu, A., Kawsar, F., Berthouze, N. & Lane, N. D. (2019). Mic2Mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems. *18th Intl Conf Inf Proc Sensor Netw*, pp. 169–180.

Maurya, A., Kumar, D. & Agarwal, R. (2018). Speaker recognition for Hindi speech signal using MFCC-GMM approach. *Procedia Computer Science*, 125, 880–887.

McFee, B., Humphrey, E. J. & Bello, J. P. (2015a). A Software Framework for Musical Data Augmentation. *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 248–254.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015b). Librosa: Audio and Music Signal Analysis in Python. *14th Python in Science Conf*, 8.

MCV. (2019). Mozilla common voice dataset.

Mello, A., de Matos, J., Stemmer, M., Britto Jr., A. & Koerich, A. (2019). A Novel Orthogonal Direction Mesh Adaptive Direct Dual Search Approach for SVM Hyperparameter Tuning. *arXiv preprint arXiv:1904.11649*.

Meng, D. & Chen, H. (2017). Magnet: a two-pronged defense against adversarial examples. *ACM SIGSAC Conf Comp and Commun Secur*, pp. 135–147.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B. & Virtanen, T. (2017). DCASE 2017 challenge setup: Tasks, datasets and baseline system. *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

Metzen, J. H., Genewein, T., Fischer, V. & Bischoff, B. (2017). On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*.

Meyer, Y. (1992). *Wavelets and Operators: Volume 1*. Cambridge university press.

Mirza, M. & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mitra, V. & Wang, C.-J. (2008). Content based audio classification: a neural network approach. *Soft Computing*, 12(7), 639–646.

Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. (2018). Spectral Normalization for Generative Adversarial Networks. *Intl Conf Learn Repres*.

Miyazaki, K., Komatsu, T., Hayashi, T., Watanabe, S., Toda, T. & Takeda, K. (2020). Weakly-supervised sound event detection with self-attention. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70.

Moosavi-Dezfooli, S.-M., Fawzi, A. & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. *IEEE Conf Comp Vis Patt Recog*, pp. 2574–2582.

Mozilla-DeepSpeech. (2017). Mozilla. Project DeepSpeech.

Mroueh, Y. & Sercu, T. (2017). Fisher GAN. *Adv Neural Inf Process Syst*, pp. 2513–2523.

Mroueh, Y., Li, C., Sercu, T., Raj, A. & Cheng, Y. (2018). Sobolev GAN. *6th Intl Conf Learn Repres*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Adv Appl Probab*, 429–443.

Mun, S., Park, S., Han, D. K. & Ko, H. (2017). Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. *Proc. DCASE*, 93–97.

Mushtaq, Z. & Su, S.-F. (2020). Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167, 107389.

Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J., Shao, L. et al. (1995). Extra-Curricular.

Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J. & Shao, L. (2019). Adversarial defense by restricting the hidden space of deep neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3385–3394.

Mydlarz, C., Salamon, J. & Bello, J. P. (2017). The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117, 207–218.

Nasiri, A. & Hu, J. (2021). SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *arXiv preprint arXiv:2103.01929*.

Nayebi, A. & Ganguli, S. (2017). Biologically inspired protection of deep networks from adversarial attacks. *arXiv Prepr arXiv:1703.09202*.

Noda, S., Mori, S., Ishibashi, F. & Itomi, K. (1987). Effect of Coils on Natural Frequencies of Stator Cores in Small Induction Motors. *IEEE Transactions on Energy Conversion*, EC-2(1), 93–99.

Palanisamy, K., Singhania, D. & Yao, A. (2020). Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*.

Palaz, D., Doss, M. M. & Collobert, R. (2015). Convolutional neural networks-based continuous speech recognition using raw speech signal. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4295–4299.

Paliwal, K. K. (1999). Decorrelated and liftered filter-bank energies for robust speech recognition. *6th European Conf Speech Comm Techn*.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. *IEEE Intl Conf Acoust, Speech and Signal Process*, pp. 5206–5210.

Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P. & Makedon, F. (2017). Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(2), 26.

Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016a). Distillation as a defense to adversarial perturbations against deep neural networks. *IEEE Symposium on Security and Privacy*, pp. 582–597.

Papernot, N. (2018). Characterizing the limits and defenses of machine learning in adversarial settings.

Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A. et al. (2016b). Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*.

Papernot, N., McDaniel, P. & Goodfellow, I. (2016c). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv Prepr arXiv:1605.07277*.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. & Swami, A. (2016d). The limitations of deep learning in adversarial settings. *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Park, H. & Yoo, C. D. (2020). CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification. *IEEE Signal Processing Letters*, 27, 411–415.

Patel, I. & Rao, Y. S. (2010). Speech recognition using hidden Markov model with MFCC-subband technique. *2010 International Conference on Recent Trends in Information, Telecommunication and Computing*, pp. 168–172.

Patidar, S. & Pachori, R. B. (2014). Classification of cardiac sound signals using constrained tunable-Q wavelet transform. *Expert Systems with Applications*, 41(16), 7161–7170.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011a). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011b). Scikit-learn: Machine learning in Python. *J Mach Learn Research*, 12, 2825–2830.

Peinado, A. M. & Segura, J. C. (2006). *Speech Recognition over Digital Channels*. John Wiley & Sons, Ltd.

Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. C. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *32nd AAAI Conf Artificial Intelligence*, pp. 3942–3951. Consulted at https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16528.

Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Allyn and Bacon Boston.

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6.

Piczak, K. J. (2015b). ESC: Dataset for environmental sound classification. *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018.

Pons, J. & Serra, X. (2019). Randomly weighted CNNs for (music) audio classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Poursaeed, O., Katsman, I., Gao, B. & Belongie, S. (2018). Generative adversarial perturbations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al. (2011). The Kaldi speech recognition toolkit. *IEEE Works Autom Speech Recog Underst*.

Pryadi, E., Gandi, K. & Kanalebe, H. Y. (2008). Speech compression using CELP speech coding technique in GSM AMR. *2008 5th IFIP International Conference on Wireless and Optical Communications Networks (WOCN'08)*, pp. 1–4.

Qian, S. (2002). *Introduction to time-frequency and wavelet transforms*. Prentice Hall PTR Upper Saddle River (NJ).

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I. & Raffel, C. (2019). Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. *Intl Conf Mach Learn*, pp. 5231–5240.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.

Radford, A., Metz, L. & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *4th Intl Conf Learn Repres*.

Radhakrishnan, R., Divakaran, A. & Smaragdis, A. (2005). Audio analysis for surveillance applications. *IEEE Workshop Appl Signal Proc Audio Acous*, pp. 158–161.

Rahman, R., Rahman, M. A., Hossain, S., Hossain, S., Akhond, M. R. & Hossain, M. I. (2021). RansomListener: Ransom Call Sound Investigation Using LSTM and CNN Architectures. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pp. 509–516.

Rauber, J., Brendel, W. & Bethge, M. (2017). Foolbox v0.8.0: A Python toolbox to benchmark the robustness of machine learning models. *CoRR*, abs/1707.04131. Consulted at http://arxiv.org/abs/1707.04131.

Rix, A. W., Beerends, J. G., Hollier, M. P. & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *IEEE Intl Conf Acoust, Speech, Signal Process*, 2, 749–752.

Rizzo, M. L. & Székely, G. J. (2016). Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1), 27–38.

Rouhani, B. D., Samragh, M., Javidi, T. & Koushanfar, F. (2017). Curtail: Characterizing and thwarting adversarial deep learning. *arXiv preprint arXiv:1709.02538*.

Roy, N., Hassanieh, H. & Roy Choudhury, R. (2017). Backdoor: Making microphones hear inaudible sounds. *15th Intl Conf Mob Sys App Serv*, pp. 2–14.

Rozsa, A., Rudd, E. M. & Boult, T. E. (2016). Adversarial diversity and hard positive generation. *Proc. IEEE Conf on Computer Vision and Pattern Recognition Workshops*, pp. 25–32.

Sabour, S., Cao, Y., Faghri, F. & Fleet, D. J. (2015). Adversarial manipulation of deep representations. *arXiv Prepr arXiv:1511.05122*.

Sailor, H. B., Agrawal, D. M. & Patil, H. A. (2017). Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification. *InterSpeech*, 8, 9.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B. & Ramabhadran, B. (2013). Deep convolutional Neural networks for LVCSR. *IEEE Intl Conf Acoust, Speech, Signal Process*, pp. 8614–8618.

Salamon, J., Jacoby, C. & Bello, J. P. (2014a, Nov.). A Dataset and Taxonomy for Urban Sound Research. *22nd ACM Intl Conf on Multimedia*.

Salamon, J. & Bello, J. P. (2015a). Feature learning with deep scattering for urban sound analysis. *23rd European Signal Processing Conference (EUSIPCO)*, pp. 724–728.

Salamon, J. & Bello, J. P. (2015b). Unsupervised feature learning for urban sound classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 171–175.

Salamon, J. & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3), 279–283.

Salamon, J., Jacoby, C. & Bello, J. P. (2014b). A dataset and taxonomy for urban sound research. *22nd ACM international conference on Multimedia*, pp. 1041–1044.

Salamon, J., MacConnell, D., Cartwright, M., Li, P. & Bello, J. P. (2017). Scaper: A library for soundscape synthesis and augmentation. *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344–348.

Salimans, T., Zhang, H., Radford, A. & Metaxas, D. (2018). Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*.

Sallo, R. A., Esmaeilpour, M. & Cardinal, P. (2021). Adversarially Training for Audio Classifiers. *International Conference on Pattern Recognition (ICPR)*, pp. 1-8, Accepted for Publication.

Samangouei, P., Kabkab, M. & Chellappa, R. (2018a). Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *Intl Conf Learn Repres*.

Samangouei, P., Kabkab, M. & Chellappa, R. (2018b). Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Saxe, A. M., McClelland, J. L. & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *2nd Intl Conf Learn Repres, ICLR*. Consulted at http://arxiv.org/abs/1312.6120.

Scardapane, S. & Uncini, A. (2017). Semi-supervised echo state networks for audio classification. *Cognitive Computation*, 9(1), 125–135.

Schönherr, L., Kohls, K., Zeiler, S., Holz, T. & Kolossa, D. (2018). Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *Netw and Distrib Syst Secur Symposium*, pp. 1-15.

Schönherr, L., Eisenhofer, T., Zeiler, S., Holz, T. & Kolossa, D. (2020). Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems. *Annual Comp Secur Appl Conf*, pp. 843–855.

Segura, J. C., Benítez, M. C., Torre, Á. d. l. & Rubio, A. J. (2002). Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR. *7th International Conference on Spoken Language Processing*.

Sengupta, S., Yasmin, G. & Ghosal, A. (2019). Speaker Recognition Using Occurrence Pattern of Speech Signal. In *Recen Trends Sign Image Proc* (pp. 207–216). Springer.

Serizel, R., Turpault, N., Eghbal-Zadeh, H. & Shah, A. P. (2018). Large-scale weakly labeled semi-supervised sound event detection in domestic environments. *arXiv preprint arXiv:1807.10501*.

Serra, J., Serra, X. & Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9), 093017.

Shah, S. K., Tariq, Z. & Lee, Y. (2018). Audio iot analytics for home automation safety. *2018 IEEE International Conference on Big Data (Big Data)*, pp. 5181–5186.

Shah, S. K., Tariq, Z. & Lee, Y. (2019). Iot based urban noise monitoring in deep learning using historical reports. *2019 IEEE International Conference on Big Data (Big Data)*, pp. 4179–4184.

Sharma, J., Granmo, O.-C. & Goodwin, M. (2019). Environment sound classification using multiple feature channels and attention based deep convolutional neural network. *arXiv preprint arXiv:1908.11219*.

Shen, J., Nguyen, P., Wu, Y., Chen, Z., Chen, M. X., Jia, Y., Kannan, A., Sainath, T., Cao, Y., Chiu, C.-C. et al. (2019). Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*.

Shi, L., Ahmad, I., He, Y. & Chang, K. (2018). Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. *Journal of Communications and Networks*, 20(5), 509–518.

Simard, P. Y., Steinkraus, D. & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *null*, pp. 958.

Simonyan, K. & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Smith, J. O. et al. (2011). *Spectral audio signal processing*. W3K.

Song, L. & Mittal, P. (2017). Inaudible voice commands. *arXiv Prepr arXiv:1708.07238*.

Song, Y., Shu, R., Kushman, N. & Ermon, S. (2018). Constructing unrestricted adversarial examples with generative models. *arXiv preprint arXiv:1805.07894*.

Spitkovsky, V. I., Alshawi, H., Jurafsky, D. & Manning, C. D. (2010). Viterbi training improves unsupervised dependency parsing. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 9–17.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R. et al. (2012). On the empirical estimation of integral probability metrics. *Electr Journal Statis*, 6, 1550–1599.

Steele, D., Krijnders, J. & Guastavino, C. (2013). The sensor city initiative: cognitive sensors for soundscape transformations. *GIS Ostrava*, 1–8.

Stephane, M. (1999). A wavelet tour of signal processing. Elsevier.

Sterne, J. (2015). Space within space: artificial reverb and the detachable echo. *Grey Room*, 110–131.

Stowell, D. & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2, e488.

Strang, G. (1993). The fundamental theorem of linear algebra. *The American Mathematical Monthly*, 100(9), 848–855.

Su, F., Yang, L., Lu, T. & Wang, G. (2011). Environmental sound classification for scene recognition using local discriminant bases and HMM. *Proceedings of the 19th ACM international conference on Multimedia*, pp. 1389–1392.

Su, Y., Zhang, K., Wang, J. & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors*, 19(7), 1733.

Sun, S., Yeh, C.-F., Hwang, M.-Y., Ostendorf, M. & Xie, L. (2018). Domain adversarial training for accented speech recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Syafria, F., Buono, A. & Silalahi, B. P. (2014). A comparison of backpropagation and LVQ: A case study of lung sound recognition. *2014 International Conference on Advanced Computer Science and Information System*, pp. 402–407.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J. & Fergus, R. (2014). Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conf Comp Vis Patt Recog*, pp. 1–9.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State Univ, Dept Mathematics and Statistics Tech Rep*, 3(05), 1–18.

Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal Statis Plann Inference*, 143(8), 1249–1272.

Szurley, J. & Kolter, J. Z. (2019). Perceptual based adversarial audio attacks. *arXiv preprint arXiv:1906.06355*.

Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans Audio, Speech, Language Process*, 19(7), 2125–2136.

Tak, R. N., Agrawal, D. M. & Patil, H. A. (2017). Novel phase encoded mel filterbank energies for environmental sound classification. *International Conference on Pattern Recognition and Machine Intelligence*, pp. 317–325.

Takahata, M., Imai, M. & Tsuji, S. (1992). Determining motion of nonrigid objects by active tubes. *1992 11th IAPR International Conference on Pattern Recognition*, 1, 647–650.

Tan, B. T., Lang, R., Schroder, H., Spray, A. & Dermody, P. (1994). Applying wavelet analysis to speech segmentation and classification. *Wavelet applications*, 2242, 750–761.

Tan, C. M. J. & Motani, M. (2020). DropNet: Reducing Neural Network Complexity via Iterative Pruning. *International Conference on Machine Learning*, pp. 9356–9366.

Tang, H., Ma, G., Chen, Y., Guo, L., Wang, W., Zeng, B. & Zhan, L. (2020). Adversarial attack on hierarchical graph pooling neural networks. *arXiv preprint arXiv:2005.11560*.

Tang, Q.-H., Liu, B.-H., Chen, Y.-Q., Zhou, X.-H. & Ding, J.-S. (2007). Application of LVQ neural network combined with the genetic algorithm in acoustic seafloor classification. *Chinese Journal of Geophysics*, 50(1), 291–298.

Taori, R., Kamsetty, A., Chu, B. & Vemuri, N. (2019). Targeted adversarial examples for black box audio systems. *IEEE Secur Priv Works*, pp. 15–20.

Thomae, C. & Dominik, A. (2016). Using deep gated RNN with a convolutional front end for end-to-end classification of heart sound. *2016 Computing in Cardiology Conference (CinC)*, pp. 625–628.

Toffa, O. K. & Mignotte, M. (2020). Environmental Sound Classification Using Local Binary Pattern and Audio Features Collaboration. *IEEE Transactions on Multimedia*, 1–1, . doi: 10.1109/TMM.2020.3035275.

Tokozume, Y. & Harada, T. (2017). Learning environmental sounds with end-to-end convolutional neural network. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2721–2725.

Tokozume, Y., Ushiku, Y. & Harada, T. (2018). Learning from Between-class Examples for Deep Sound Recognition. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D. & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

Tsalera, E., Papadakis, A. & Samarakou, M. (2020). Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm. *Energy Reports*, 6, 223–230.

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5), 293–302.

Uesato, J., O'donoghue, B., Kohli, P. & Oord, A. (2018). Adversarial risk and the dangers of evaluating against weak attacks. *International Conference on Machine Learning*, pp. 5025–5034.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. CoRR abs/1607.08022 (2016).

Umapathy, K., Krishnan, S. & Rao, R. K. (2007). Audio signal feature extraction and classification using local discriminant bases. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), 1236–1246.

Vacher, M., Serignat, J.-F. & Chaillol, S. (2007). Sound classification in a smart room environment: an approach using GMM and HMM methods. *The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD 2007), Publishing House of the Romanian Academy (Bucharest)*, 1, 135–146.

Vaizman, Y., McFee, B. & Lanckriet, G. (2014). Codebook-based audio feature representation for music information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(10), 1483–1493.

Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F. & Sarti, A. (2007). Scream and gunshot detection and localization for audio-surveillance systems. *IEEE Conf Adv Video Sign Based Surv*, pp. 21–26.

Valero, X. & Alías, F. (2012a). Classification of audio scenes using narrow-band autocorrelation features. *2012 Proceedings of the 20th European signal processing conference (EUSIPCO)*.

Valero, X. & Alías, F. (2012b). Gammatone wavelet features for sound classification in surveillance applications. *2012 Proceedings of the 20th European Signal Processing*

Conference (EUSIPCO), pp. 1658–1662.

Van Fleet, P. J. (2011). *Discrete wavelet transformations: An elementary approach with applications*. John Wiley & Sons.

Van Loan, C. F. & Golub, G. H. (1983). *Matrix computations*. Johns Hopkins University Press.

Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *25th Intl Conf Mach Learn*, pp. 1096–1103.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans Speech Audio Process*, 7(2), 126–137.

Wall, M. E., Rechtsteiner, A. & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis* (pp. 91–109). Springer.

Wang, H., Zou, Y., Chong, D. & Wang, W. (2019). Environmental sound classification with parallel temporal-spectral attention. *arXiv preprint arXiv:1912.06808*.

Wang, H., Zou, Y., Chong, D. & Wang, W. (2020). Environmental sound classification with parallel temporal-spectral attention. *21st Annual Conf Intl Speech Comm Assoc*.

Wang, J.-C., Wang, J.-F., He, K. W. & Hsu, C.-S. (2006). Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor. *The 2006 IEEE international joint conference on neural network proceedings*, pp. 1731–1735.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Watkins, D. S. (2007). *The matrix eigenvalue problem: GR and Krylov subspace methods*. SIAM.

Weiping, Z., Jiantao, Y., Xiaotao, X., Xiangtao, L. & Shaohu, P. (2017). Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. *2017 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J. & Daniel, L. (2018). Soundnet: Learning sound representations from unlabeled video. *6th Intl Conf Learn Repres*.

Wierstra, D., Schaul, T., Peters, J. & Schmidhuber, J. (2008). Natural evolution strategies. *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational*

Intelligence), pp. 3381–3387.

Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., Xu, P. & Fung, P. (2020). Meta-transfer learning for code-switched speech recognition. *arXiv preprint arXiv:2004.14228*.

Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559*.

Xiao, H., Xiao, H. & Eckert, C. (2012). Adversarial Label Flips Attack on Support Vector Machines. *ECAI*, pp. 870–875.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L. & Yuille, A. (2017). Adversarial examples for semantic segmentation and object detection. *IEEE Intl Conf Comp Vis*, pp. 1369–1378.

Xie, C., Zhang, Z., Wang, J., Zhou, Y., Ren, Z. & Yuille, A. (2018). Improving Transferability of Adversarial Examples with Input Diversity. *arXiv Prepr arXiv:1803.06978*.

Xu, M., Xu, C., Duan, L., Jin, J. S. & Luo, S. (2008). Audio keywords generation for sports video analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 4(2), 11.

Yakura, H. & Sakuma, J. (2018). Robust audio adversarial example for a physical attack. *28th Intl J Conf Artif Intell*, pp. 5334-5341.

Ye, J., Kobayashi, T. & Murakawa, M. (2017). Urban sound event classification based on local and global features aggregation. *Applied Acoustics*, 117, 246–256.

Yoma, N. B. & Villar, M. (2002). Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm. *IEEE Transactions on Speech and Audio Processing*, 10(3), 158–166.

Young, R. K. (2012). *Wavelet theory and its applications*. Springer Science & Business Media.

Yu, G. & Slotine, J.-J. (2008). Audio classification from time-frequency texture. *arXiv Prepr arXiv:0809.4501*.

Yu, G., Mallat, S. & Bacry, E. (2008). Audio denoising by time-frequency block thresholding. *IEEE Trans Sign Proc*, 56(5), 1830–1839.

Yuan, X., He, P., Zhu, Q. & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9), 2805–2824.

Zamil, M. G. A., Samarah, S., Rawashdeh, M., Karime, A. & Hossain, M. S. (2019). Multimedia-oriented action recognition in Smart City-based IoT using multilayer perceptron. *Multimedia Tools and Applications*, 78(21), 30315–30329.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z. & Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Zhang, B., Quan, C. & Ren, F. (2016). Study on CNN in the recognition of emotion in audio and images. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–5.

Zhang, J., Zhang, B. & Zhang, B. (2019a). Defending adversarial attacks on cloud-aided automatic speech recognition systems. *Proceedings of the Seventh International Workshop on Security in Cloud Computing*, pp. 23–31.

Zhang, X., Zou, Y. & Shi, W. (2017). Dilated convolution neural network with LeakyReLU for environmental sound classification. *2017 22nd International Conference on Digital Signal Processing (DSP)*, pp. 1–5.

Zhang, X., Zou, Y. & Wang, W. (2018a). Ld-CNN: a lightweight dilated convolutional neural network for environmental sound classification. *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 373–378.

Zhang, Z., Xu, S., Cao, S. & Zhang, S. (2018b). Deep convolutional neural network with mixup for environmental sound classification. *Chinese conference on pattern recognition and computer vision (prcv)*, pp. 356–367.

Zhang, Z., Xu, S., Zhang, S., Qiao, T. & Cao, S. (2019b). Learning attentive representations for environmental sound classification. *IEEE Access*, 7, 130327–130339.

Zhang, Z. & Schuller, B. (2012). Semi-supervised learning helps in sound event classification. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 333–336.

Zhu, B., Xu, K., Wang, D., Zhang, L., Li, B. & Peng, Y. (2018a). Environmental Sound Classification Based on Multi-temporal Resolution Convolutional Neural Network Combining with Multi-level Features. In *Advances in Multimedia Information Processing* (pp. 528–537). Cham: Springer International Publishing.

Zhu, B., Xu, K., Wang, D., Zhang, L., Li, B. & Peng, Y. (2018b). Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features. *Pacific Rim Conference on Multimedia*, pp. 528–537.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.

Zhu, Q. & Alwan, A. (2001). An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, 1, 113–116.

Zhu, X., Liu, Y., Qin, Z. & Li, J. (2017b). Data augmentation in emotion classification using generative adversarial networks. *arXiv preprint arXiv:1711.00648*.

Zhu, X., Liu, Y., Li, J., Wan, T. & Qin, Z. (2018c). Emotion Classification with Data Augmentation Using Generative Adversarial Networks. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 349–360.