EvolMusic: Towards Musical Adversarial Examples for Black-Box Attacks on Speech-To-Text

Mariele Motta
neurocat GmbH
Berlin, Germany
mm@neurocat.ai

Tanja Hagemann
Telekom Innovation Laboratories
Berlin, Germany
tanja.hagemann@telekom.de

Sebastian Fischer
Telekom Innovation Laboratories
Berlin, Germany
sebastian-fischer@telekom.de

Felix Assion
neurocat GmbH
Berlin, Germany
fa@neurocat.ai

ABSTRACT

Automatic Speech Recognition (ASR) has undergone substantial improvements since the incorporation of deep learning. However, the vulnerability of neural networks to imperceptible adversarial perturbations exposes ASR-based devices to potentially serious threats. So far, imperceptibility of audio adversarial examples has been associated with small, or inaudible perturbations. In this paper, we expand the domain of viable audio adversarial examples to include audible, but inconspicuous adversarial perturbations. We present `EvolMusic`, the first targeted adversarial attack based on musical note-sequences. Our musical perturbations are generated via an adaptive evolutionary approach in a black-box setting. We evaluate our attack against DeepSpeech v0.9.1 using the Fluent Speech Commands dataset.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; **Speech recognition**; **Machine learning**.

KEYWORDS

adversarial attack, black-box, deep learning, genetic algorithm, speech recognition

ACM Reference Format:
Mariele Motta, Tanja Hagemann, Sebastian Fischer, and Felix Assion. 2021. EvolMusic: Towards Musical Adversarial Examples for Black-Box Attacks on Speech-To-Text. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459488

1 INTRODUCTION

Automatic Speech Recognition (ASR) is an increasingly pervasive technology with security-critical applications such as in-car navigation systems, smart home devices, and telephone assistance lines. The incorporation of deep learning into ASR systems introduces a vulnerability to adversarial examples – inputs crafted with the purpose of misleading the system while going unnoticed by humans [1–3]. Several techniques for crafting such malicious inputs have been developed in both, white and black-box settings, with most of the applications in the computer vision domain [4]. Research in the adversarial robustness of ASR systems has experienced recent progress, but it is still incipient [5–7].

With this work, we demonstrate that it is possible to craft musical adversarial perturbations which can change an input classification to a chosen target. We consider a black-box setting and generate musical note-sequences via an adaptive evolutionary approach. Such musical perturbations are audible, but may not be perceived as harmful and can potentially go unnoticed if embedded into some musical background. We consider DeepSpeech v0.9.1 as our target model and evaluate `EvolMusic` using common audio commands.

2 THREAT MODEL

The goal of our attack is to change the transcription of an audio input to a target prediction by adding musical perturbations to it. In analogy with evolutionary biology, a musical note-sequence corresponds to the genotype and the adversarial example, i.e. the note-sequence in addition to the original audio, corresponds to the phenotype. The fitness score of an adversarial example is given by the edit distance between its prediction by the model and the target. We denote by D^g the vector of scores of generation g.

Figure 1 illustrates how `EvolMusic` works in a simplified setting, with different colors indicating different notes. After creating the initial population with N members, we iterate through steps $1-5$ until the fittest member's prediction matches the target, or until a chosen maximum number of generations is reached. At each step, parents are selected with probabilities $p^g = sigmoid(D^g)$. We make the following design choices to help to escape local minima:

1) Adaptive probabilities for adding mutants [1] as parents p^g_{mutant}, and for mutation of the children [2] p^g_{mutate}:

$$p^g_{mutant} = p^g_{mutate} = \frac{\alpha}{\text{sig}\,(N \cdot \Delta)}, \tag{1}$$

[1] Mutants are random note-sequences.
[2] The mutation of a child corresponds to the addition of a random value to each note following a normal distribution $\mathcal{N}(0, mut_{std})$, with mut_{std} a hyperparameter to be tuned.

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459488

where $\Delta = \sigma(D^g)/\mu(D^g)$ is the ratio between the standard deviation and the average of the scores vector of the elite members [3] D^g at generation g and α is the minimum probability value, which can be adjusted for each mutation operation. $sig(\cdot)$ denotes a sigmoid-like function, $sig(x) = 1/(c + exp(-x))$, and the constant c is adjusted to control the maximum value of p^g_{mutant} and p^g_{mutate}.

2) Different types of crossover, illustrated in Figure 2: We consider both a standard and a complementary split point – the latter switches the order of the notes. We also allow for a piece-wise crossover, which creates children based on the parts of the parents which produce a better target match.

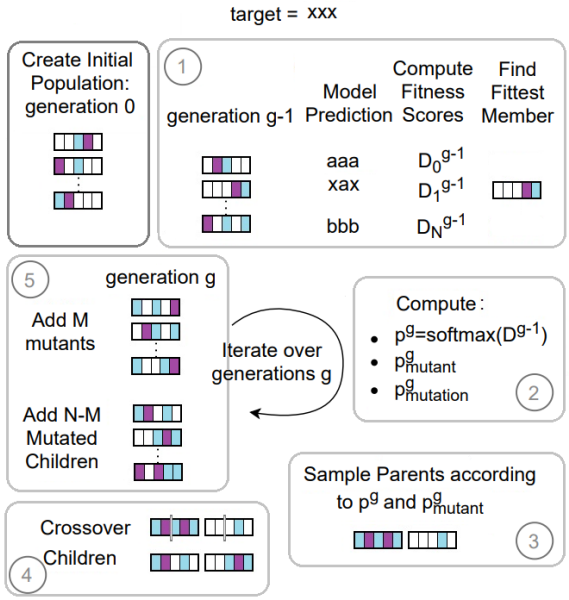

Figure 1: `EvolMusic`: population size N, with M mutants added at each generation.

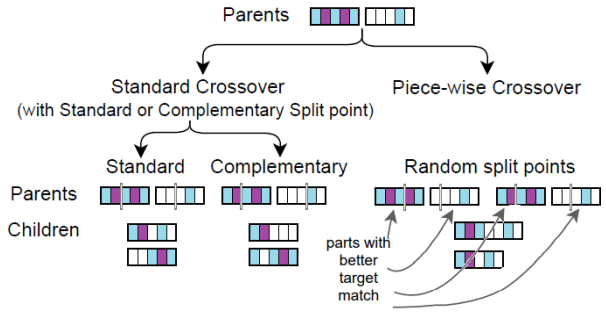

Figure 2: Crossover Types: The dark, downward arrows indicate a coin toss to select the type of operation.

3 EXPERIMENTS

We benchmark our attack on randomly selected audio files from the Fluent Speech Commands dataset [8] and use a maximum of 3000 iterations to run the attack. We consider two sets of target predictions: 1) The transcriptions of the Google Speech Commands dataset [9], and 2) The top 20 commands uttered to Google Home [10]. We report the target success rate and the average target similarity, defined as $1 -$ *the average normalized edit distance*.

[3] Elite members are the best N members of the generations g and $g-1$.

Table 1: Results for all ten targets obtained from the Google Speech Commands dataset

Original	Target	Prediction	Score*
switch off the washroom lights	yes	yes	0.0
make it quieter	down	down	0.0
switch language	stop	stop	0.0
switch off the washroom lights	go	go	0.0
play the music	no	no	0.0
make it louder	on	on	0.0
bathroom lights on	left	let	0.25
I need volume	off	of	0.33
bedroom heat down	right	eighteen	0.5
turn the kitchen temperature down	up	upon	0.5

* normalized edit distance between target and prediction

Results for targets from the Google Speech Commands dataset are summarized in Table 1. We obtain an average target success rate of 60% and an average target similarity of 84% for this dataset.

On targets from the top 20 Google Home commands, our attack achieves a target success rate of 15% and an average target similarity of 63%. The average word error rate between the original transcriptions and the predicted targets is 105%.

4 CONCLUSION

With `EvolMusic` we have demonstrated that it is possible to generate targeted attacks in a black-box setting with sequences of musical notes. To the best of our knowledge, this is the first targeted adversarial attack based on musical perturbations.

REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.
[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.
[3] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017.
[4] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, and Mani B. Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. *CoRR*, abs/1805.11090, 2018.
[5] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *Proceedings of the 25th USENIX Conference on Security Symposium*, SEC'16, page 513–530, USA, 2016. USENIX Association.
[6] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.
[7] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *CoRR*, abs/1808.05665, 2018.
[8] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding, 2019.
[9] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209, 2018.
[10] Frank Bentley, Luvogt Chris, Silverman Max, Wirasinghe Rushani, White Brooke, and Lottrjdge Danielle. Understanding the longterm use of smart speaker assistants. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 3*, 2018.