Black-Box Universal Adversarial Attack on Automatic Speech Recognition Systems for Maritime Radio Communication Using Evolutionary Strategies

Aliza Katharina Reif
aliza.reif@dlr.de
AI Safety and Security
German Aerospace Center (DLR)
St. Augustin, Germany

Lorenzo Bonasera
lorenzo.bonasera@dlr.de
AI Safety and Security
German Aerospace Center (DLR)
St. Augustin, Germany

Stjepan Picek
stjepan.picek@ru.nl
Radboud University
Nijmegen, Netherlands

Oscar Hernán Ramírez-Agudelo
oscar.ramirezagudelo@dlr.de
AI Safety and Security
German Aerospace Center (DLR)
St. Augustin, Germany

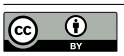
Michael Karl
michael.karl@dlr.de
AI Safety and Security
German Aerospace Center (DLR)
St. Augustin, Germany

Abstract

This paper studies the design, implementation, and evaluation of a new universal adversarial attack targeting automatic speech recognition systems in a black-box setting. A genetic algorithm optimizes universal perturbations consisting of short noise bursts that cause mistranscriptions by balancing text similarity (character error rate) and perceptual audio similarity (Mel energy distance) to keep the noise minimally intrusive. Experiments are conducted on the models Wav2Vec 2.0 and OpenAI's Whisper to investigate the attack's performance under varying parameters such as noise volumes, number of audio files in the training set, and for the standard English Librispeech dataset, as well as a synthetic maritime dataset that contains more homogeneous data. We expose vulnerabilities in state-of-the-art ASR systems and the risks of attacks on safety-critical applications, such as maritime radio communication. We demonstrate that our attack is highly successful, and even an attack trained on a single input works universally. Whisper proves to be more robust against these attacks. We find that universal perturbations generalize better when trained on data more similar to the test set. A semantic defense is developed that presents a novel way to detect the attack. To our knowledge, our work represents the first universal black-box attack against ASR models.

CCS Concepts

• **Security and privacy → Domain-specific security and privacy architectures**; • **Computing methodologies → Speech recognition**; • **Theory of computation → Genetic programming**.



AISec '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1895-3/2025/10
https://doi.org/10.1145/3733799.3762974

Keywords

Adversarial Attack, Universal Attack, Automatic Speech Recognition, Maritime Radio Communication, Genetic Algorithms

ACM Reference Format:
Aliza Katharina Reif, Lorenzo Bonasera, Stjepan Picek, Oscar Hernán Ramírez-Agudelo, and Michael Karl. 2025. Black-Box Universal Adversarial Attack on Automatic Speech Recognition Systems for Maritime Radio Communication Using Evolutionary Strategies. In *Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3733799.3762974

1 Introduction

Reliable maritime radio communication is crucial to the safety of operations at sea. To ensure clarity, precision, and fast reaction times, the International Maritime Organization (IMO) developed the Standard Maritime Communication Phrases (SMCP) in 2001 [11]. These phrases standardize Very High Frequency (VHF) radio communication between ships, coastal stations, and on board, in particular during emergency situations. The SMCP recognizes English as the primary language of maritime communication and is a simplified version of nautical English. Thereby, it reduces ambiguity and improves understanding across language barriers in an international environment [16, 4]. Maritime radio communication commonly relies on VHF bands, which range from 30 to 300 MHz, due to their reliable coverage. Since VHF signals propagate along line-of-sight, the communication range increases with the elevation of the transmitting antenna [12].

In recent years, Automatic Speech Recognition (ASR) systems have been shown to be useful for transcribing SMCP radio transmissions in real time to enhance maritime safety [17, 11, 15]. These transcripts can then be analyzed by Large Language Models (LLMs) to extract structured information from the standardized format with the goal of supporting decision-making and allowing automated emergency assessments. The distinctly structured nature of SMCP radio communication makes this application area well-suited for information extraction with the help of LLMs, such as transcriptions of spoken reports to standardized tabular formats [11]. However, this pipeline from speech to structured output critically relies on

the accuracy of the transcriptions by the ASR system. Both human-made noise (for example, varying languages or overlapping speech) and environmental noise (for example, background noises or static radio noise) can negatively affect the accuracy [11]. Additionally, intentional manipulations of the audio inputs can degrade the performance of ASR systems, and all these factors would also cause downstream systems to fail because of incorrect transcriptions. For many commercial ASR systems, the model parameters and architecture are not publicly known, which can mitigate white-box attacks but still makes the model vulnerable to attacks in a black-box setting. Recent research has shown that audio perturbations crafted from adversarial black-box attacks can cause ASR systems to mistranscribe audio inputs without needing any knowledge of the model's internal workings [13].

Previous black-box attacks have mainly focused on generating input-specific individual perturbations that are customized to deceive the ASR system on one specific target audio [2]. Although such attacks have proven effective in controlled environments for demonstrating feasibility [13], they are not practical for real-world maritime communication. In operational settings, adversaries typically lack prior knowledge of the transmitted messages, and individual black-box attacks are infeasible in real time due to the high number of queries required to generate adversarial perturbations [22, 23]. A more powerful and realistic threat model is the universal adversarial perturbation, which consists of a single noise vector that, when added to any new audio input, consistently induces mistranscriptions. Designing such a perturbation without access to gradients or training data remains a significant challenge.

This paper addresses this gap by introducing a novel universal black-box adversarial attack on audio data. The attack uses a genetic algorithm to evolve a fixed noise pattern that can be added to any new audio sample, including examples of maritime radio communication phrases. The universal perturbation then increases the errors in the audio transcription. The genetic algorithm iteratively optimizes this universal perturbation over generations based on queried model outputs, without requiring access to model architecture, gradients, or parameters. The noise is generated to be phoneme-aware by exploiting the known weaknesses of ASR systems in recognizing obstruent phonemes [19]. This helps to keep the noise less noticeable while increasing the attack success rate. To keep the noise below a threshold where it interferes with a human's ability to correctly understand what is said, the perturbation's energy compared to the original audio is measured, and a penalty is added if the threshold is passed. Furthermore, to mitigate this threat, we propose a text-based detection method that compares semantic patterns in adversarial transcriptions to correct transcriptions.

In summary, this paper explores the vulnerabilities of maritime radio communication to black-box adversarial attacks. The main contributions are:

- Vulnerability assessment of the audio transcription process by analyzing transcription performance under random noise conditions, with a focus on character and word error rates.
- Replication of individual black-box attack with a customized genetic algorithm that uses phoneme-aware noise generation to target known ASR weaknesses when transcribing obstruent sounds.
- Design of a novel universal black-box attack using the same phoneme-aware genetic algorithm, but to create a single input-agnostic universal noise vector that significantly increases character and word error rates.
- Design of a text-based detection method for the attack that applies knowledge of semantic differences between manipulated and non-manipulated transcripts to create a lightweight classifier.

The source code can be found here: https://github.com/Annilophir/universal_black_box_audio_attack.

2 Background

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) systems need to be able to make sense of human speech, both in the syntactic sense of acoustically understanding the words being said and the semantic context of the spoken phrases [14]. To achieve high performance on unknown audio input, these systems require extremely large training datasets to generalize well [19]. ASR systems follow the same overall structure: A speech utterance is captured and given as input to a decoder model that can also use an acoustic language model or a lexicon to hypothesize about phonemes [14]. Phonemes are the smallest unit of speech that can be distinguished based on their meaning, and they can be differentiated into vowels and consonants, with vowels often being the point of emphasis in speech and consonants often being shorter [3, 19], or into sonorants and obstruents [19]. Obstruents are sounds formed by obstructing airflow, while sonorants are formed via continuous airflow and therefore resonate [19]. It is often a challenge for ASR systems to properly recognize the short, sharp sounds of obstruent consonants, like /p,t,k/ [19]. It has been shown that ASR models are particularly sensitive to adversarial phonemes, i.e., adversarial perturbations that mimic the sound of individual phonemes [30, 25].

2.2 Adversarial Attacks on ASR Systems

Adversarial attacks on ASR systems can be classified based on adversarial knowledge or perturbation scope [2]. White-box adversarial attacks require the attacker to have knowledge of the parameters and functions of the model. In contrast, black-box adversarial attacks operate without assumptions of how the model works, as they only consider inputs and corresponding outputs, but not the inner workings of the model in between [13, 29]. This makes black-box attacks more dangerous to real-world systems because they can be applied to any system without prior knowledge requirements [13]. However, black-box attacks are also more difficult to implement [31], and often rely on a large number of queries for each perturbed input. The perturbation scope defines on which inputs an attack is effectively applied. An individual perturbation is crafted for one specific input audio and causes that input to be mistranscribed, while a universal perturbation is created to result in mistranscriptions of any input in a dataset. Because of this, once a universal perturbation is generated, it can be applied to any input in real-time [2, 29]. The majority of existing adversarial audio attacks [5, 21, 27, 28] are white-box, individual, and targeted [2, 22]. To our knowledge, no other works consider our setup of a black-box, universal, untargeted attack. Because of this, many of these

attacks cannot be applied in the real world. Indeed, they cannot be generated in real-time, and the assumption of a white-box model is not realistic for most commercial ASR systems [27]. In settings where a stream of previously unknown audio input is directly fed into an ASR system, which has the task of transcribing the audio in real-time, these attacks cannot be effectively applied.

Evolutionary strategies have also been used to generate individual adversarial attacks before. In [13], Khare et al. applied multi-objective evolutionary optimization to successfully generate targeted and untargeted black-box individual adversarial examples on audio data with a genetic algorithm. However, finding an adversarial perturbation requires a lot of queries, making it inefficient for real-time scenarios [22]. Other evolutionary algorithms have also been previously used for individual attacks [24, 7, 31].

In [18], Neekhara et al. introduced universal audio perturbations by applying a single perturbation to multiple audio inputs and achieved mistranscriptions for all of them. In the presented white-box attack, the character error rate is optimized to be as high as possible over a dataset of input examples, resulting in the generation of a single perturbation vector to cause errors in transcriptions in the majority of audio input samples to which it is added. Universal black-box adversarial attacks are much less common [2, 22]. The ones that exist are based on transferable feature extraction from white-box models, like [8], [10], and [22].

These examples show that a universal black-box method has great potential to be effective in real-time by removing the disadvantage of individual black-box attacks requiring too many queries, and because of that, too much time to be generated to be efficiently applied in real-world cases. While the number of queries is unlikely to be reduced through this method, the resulting perturbation can afterwards be applied to as many audio inputs as necessary, without needing to be recomputed, and while showing the desired effect of causing mistranscriptions.

2.3 Evolutionary Algorithms

Evolutionary Algorithms (EAs) are a class of optimization techniques inspired by natural selection and genetic evolution in biology, and the principle of "survival of the fittest" [26]. Potential solutions are called chromosomes, which contain genes, the features of a solution. A common setup evolves this population of chromosomes over subsequent generations through selection, crossover, and mutation, inspired by biological reproduction [26]. A fitness function assigns a value of how good each chromosome is as a solution to the optimization problem [26]. A chromosome dominates another chromosome if it has a strictly higher fitness value [13]. A subset of dominant chromosomes is selected to form the elite and passed into the next generation unchanged. New chromosomes for the next generation are created by recombining elite chromosomes using crossover: two elite parents reproduce by splitting their genes and using part of each parent's genes to form a new solution to explore new regions in the solution space. Mutations are applied to some new chromosomes to increase the diversity of solutions and escape local minima by randomly altering part of the genotype of some chromosomes [26]. The process is repeated over generations to find the best solution [26]. The algorithm cannot guarantee finding a global optimum as it is only heuristic, and solutions can vary widely between runs [6].

3 Threat Model

The goal of the proposed universal black-box attack is to evolve one untargeted adversarial perturbation that effectively causes mistranscriptions across a dataset. The attack has several advantages compared to previous attacks: Since the attack is black-box, it does not require any knowledge of the model and can be applied to any model. A disadvantage of black-box attacks is that they are often reliant on a large number of queries. While our attack does need many queries to find the perturbation, it is a universal attack and therefore needs only one query at test time when applying the attack on previously unseen data.

The attack is done using black-box evolutionary strategies, requiring no prior knowledge of the parameters or functions of the model and making no assumptions about the model's architecture. The model is queried with input samples, and the model's output transcriptions are evaluated while balancing between two objectives: minimizing the text similarity between the original text and the transcribed text, and maximizing the acoustic similarity between the original audio and the manipulated audio with the added perturbation.

Ideally, a universal attack is length-agnostic and can therefore be applied to audio of any length. To achieve this, the proposed attack uses a perturbation of a pre-defined short length, here, 4-second blocks, which was experimentally found to bring the highest performance while beig computationally efficient, and repeats and crops this block until the length matches the original audio's length.

During each iteration, the candidate perturbation δ is evaluated on a set of speech samples $X = \{x_1, ..., x_n\}$, layered over the original audio signals with an additive function, and optimized to maximize the transcription error while minimizing the audible distortion, to find the optimal perturbation δ^*. The two opposing optimization goals are weighted against each other in the fitness function

$$\delta^* = \arg\max_{\delta} \left[\sum_{i=0}^{N} \lambda_1 CER\big(t(x_i+\delta), t(x_i)\big) - \lambda_2 \mathcal{L}_{mask}\big(x_i+\delta, x_i\big) \right], \quad (1)$$

where $\lambda_1 \geq 0$ represents the weight of the text dissimilarity score, which is set to 1, and $\lambda_2 \geq 0$ represents the weight of the audio similarity penalty score, which is set to be 10. The values were selected through preliminary experiments.

3.1 Text Dissimilarity

Text dissimilarity is measured by the Character Error Rate (CER) between the manipulated hypothesis transcript $t(x_i + \delta)$ and the correct reference transcript $t(x_i)$. The CER is the character-level Levenshtein distance and alignment between the two transcripts, also known as the edit distance, tracking total hits, substitutions, deletions, and insertions.

$$CER = \frac{\text{EditDistance}(x, y)}{\text{length}(x)}. \quad (2)$$

When testing the effectiveness of the attack, the word error rate (WER) is also evaluated. The WER is the word-level Levenshtein

distance between the correct reference and the hypothesized transcript; unlike the CER, which looks at individual characters, the WER regards each word as a character in the same equation.

3.2 Audio Similarity

Audio similarity is measured by comparing the perturbation δ and the original audio x_i, and constraining the noise to stay below a scaled threshold of the original audio's energy in the Mel domain by introducing a penalty to the fitness function. This measurement is therefore used as a regularizer that encourages the optimizer to prevent the noise from becoming too loud while not penalizing solutions with noise below the threshold. The Mel spectrogram representations of both audio files are calculated as $Mel(x)$ and $Mel(\delta)$ using a short-time Fourier transform followed by a Mel filter. The masking threshold α is set to the acceptable maximum noise relative to the original audio, and a violation occurs if $\text{mean}(Mel(\delta)) > \alpha \cdot \text{mean}(Mel(x))$. The loss is then computed as:

$$\mathcal{L}_{mask} = max\big(0, \text{mean}(Mel(\delta)) - \alpha \cdot \text{mean}(Mel(x))\big), \quad (3)$$

where $\alpha > 0$ is the factor of how much energy the noise is allowed to have on average compared to the original audio. For example, $\alpha = 0.5$ means that the noise is penalized if it has more energy than 50% of the energy of the original audio.

3.3 Problem Formulation and Algorithm

The implemented genetic algorithm aims to generate a universal adversarial perturbation in the form of a single noise vector that, when added to an arbitrary audio sample, reduces the transcription performance of an automatic speech recognition system while maintaining similarity to the original audio.

The genetic algorithm starts by initializing a population of random noise vectors at a sample rate 16000 Hz, which is the standard for Wav2Vec 2.0 and Whisper. The noise is initialized as a compressed and sparse vector, which can later be expanded by repeating each value according to the compression factor to match the full length of the perturbation. This approach reduces the dimensionality of the noise and lowers the number of genes required by the genetic algorithm, thereby accelerating its convergence. To mimic obstruent phonemes rather than a continuous background signal, dropout is applied to produce noise in bursts. The amplitude of the noise is sampled from a uniform distribution.

The population evolves over generations by comparing the summed fitness scores of each noise vector on a training set of audio inputs, as defined in Eq. (1). A higher fitness score denotes a larger transcription error, while noise that exceeds a predefined amplitude threshold incurs a penalty that lowers the fitness. The top-ranked elite vectors are copied unchanged into the next generation and serve as parents for reproduction. During crossover, segments from two randomly chosen elite parents are combined to create new individuals. Offspring are then mutated with a specified probability by injecting short noise bursts of variable length and amplitude into the perturbation vector. After a fixed number of generations sufficient for convergence, the algorithm returns the best noise vector, which is subsequently evaluated on unseen audio samples to assess the universality of the attack. The values are provided for each experiment in Section 4.

The attack is done on subsets of varying length of the Librispeech dataset and a custom synthetic standard maritime communication phrase (SMCP) dataset, which contains audio samples that are more similar to each other in terms of part-of-speech tagging and Mel spectrogram similarity than the standard English Librispeech dataset. The datasets are further explored in Appendix A.

Wav2Vec 2.0 is used as the baseline model for experiments due to its high accuracy and speed [1]. For comparison, OpenAI's state-of-the-art Whisper base model is also used, which has a higher accuracy and is more robust to noise, but slower [9]. Whisper is also a non-deterministic model, so giving it the same input twice does not always result in the same output transcript [9].

3.4 Semantic Defense

The proposed universal black-box attack produces adversarial outputs that not only present a noticeable pattern within the waveform and Mel spectrogram but also show linguistic patterns that deviate from clean transcripts. While defense strategies aiming at the detection of an attack presence that focus on waveforms or Mel spectrograms can be highly successful, they are also computationally expensive and can be difficult to interpret depending on the level of noise. A defense strategy with the goal of instead detecting linguistic deviations from non-manipulated transcripts has the inherent advantage that it can work on a dataset built purely from manipulated versus non-manipulated strings of written text, which is computationally less expensive and highly interpretable.

Transcripts are generated from successful runs of the genetic algorithm, using varying noise lengths, compression rates, and noise thresholds. The data is then split into training and test sets. After generating the transcript data with and without noise, the data are given a binary label indicating whether the data is clean (0) or adversarially perturbed (1). A batch size of 16 is used. For the classification task, a convolutional neural network (CNN) classifier with 3 convolutional layers, batchnorms, a ReLU activation function, a max pooling layer, dropout, and two linear layers at the end is used. The classifier takes embedded text as input and extracts contextual representations of the entire input sequence. The model is trained with a binary cross-entropy loss with logits and an Adam optimizer with a learning rate of 0.001 for 5 epochs and finally evaluated on the previously unseen test data.

4 Experimental Results

4.1 Baseline Evaluation on Clean Data

The transcription performance of the Wav2Vec model is examined on clean audio data from the Librispeech dataset. As shown in Figure 1, most of the samples are transcribed with little error. The average CER on clean data is 0.027, the average WER is 0.071. From this, it can be concluded that the ASR system is generally robust under benign conditions with only a few errors, and it is possible to establish that an attack can be deemed successful if its attack success rates are significantly higher than these baselines.

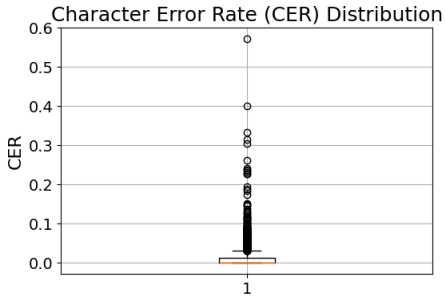


Figure 1: Boxplot showing the variance in CER values of clean audio samples from the Librispeech dataset. Most samples contain very few errors, with some outliers.

4.2 Baseline Evaluation on Random Noise

To demonstrate the effectiveness of the genetic algorithm in learning to optimize adversarial noise that increases the attack success rate, a baseline comparison is conducted that uses random noise drawn from the same distribution as the initial population of noise in the genetic algorithm. This demonstrates that the improvements achieved by the genetic algorithm are not the result of arbitrary noise but of learned perturbations that actively exploit weaknesses in the transcription process.

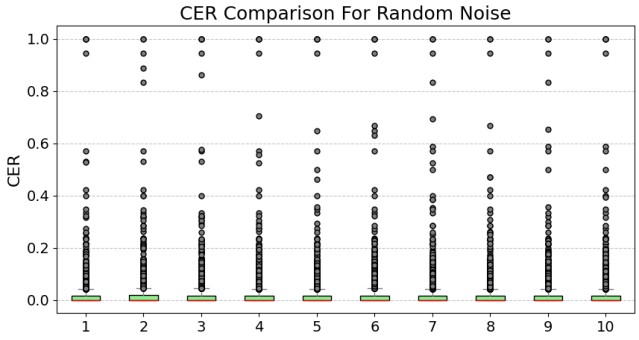


Figure 2: Boxplots showing the CER value distribution for 10 random perturbations on the Librispeech dataset. There are no significant differences between the perturbations. It can be seen that most samples have a very low error rate with some outliers.

The CER and WER do not deviate significantly from the clean baseline, which can be seen in the boxplots in Figure 2, showing that for all random noise vectors that were tested, the majority of samples were correctly transcribed with only a few outliers that contained errors.

4.3 Replication of Individual Black-box Attack

The research by Khare et al. [13] has already demonstrated that it is possible to create an adversarial audio attack using a genetic multi-objective optimization algorithm and cause an audio input to be mistranscribed by adding a noise vector that is specifically optimized for that specific individual audio signal. Our goal now is to show that the threshold-based genetic algorithm that enforces a penalty if the noise is too loud and uses noise bursts to mimic obstruent phonemes is also successful as an individual attack.

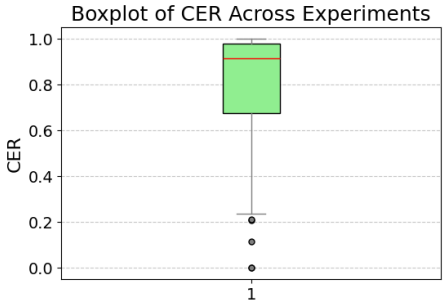


Figure 3: Boxplot showing the variance in CER values of individually attacked audio samples. Most samples show a high error rate, indicating that the attack is very successful.

In Figure 3, the character error rates per individual attack are, on average, very high, and their variance is low, with only a few outliers in which the attack did not work as well.

4.4 Universal Attack

Applying the same genetic algorithm, a universal attack is created that takes multiple audio signals as input and optimizes an adversarial perturbation on all of them at the same time. The attack is considered successful if the CER or WER is greater than 20%, respectively. The threshold was chosen because differences between the reference transcript and the mistranscribed hypothesis become noticeable at this value. The attack is overall highly successful. Each experiment is repeated 10 times over different audio inputs to account for non-determinism.

Experiments on the noise threshold (NT)				
NT	avg CER ± std	avg WER	CER ASR	WER ASR
0.2	0.054 ± 0.001	0.116	0.170	0.405
0.3	0.197 ± 0.009	0.312	0.564	0.735
0.4	0.244 ± 0.010	0.369	0.562	0.741
0.5	0.365 ± 0.009	0.505	0.735	0.860
0.6	0.398 ± 0.009	0.538	0.704	0.833
0.8	0.467 ± 0.009	0.608	0.769	0.879
1	0.532 ± 0.012	0.667	0.851	0.926

Table 1: Experiments on the influence of the noise threshold (NT). Unchanged parameters: compression factor = 10**, initial noise epsilon** = 0.02**, noise length** = $4 \cdot 16,000$**, mutation rate** = 0.2**, elite fraction** = 0.2**, population size** = 100**, generations** = 30**, number of samples in training** = 10**, dropout probability to create noise bursts** = 0.95**. CER ASR and WER ASR indicate the highest CER and WER attack success rate achieved over all experiments, respectively.**

It can be clearly observed in Figure 4 and Table 1 that a higher noise threshold results in a more successful attack. This is logical

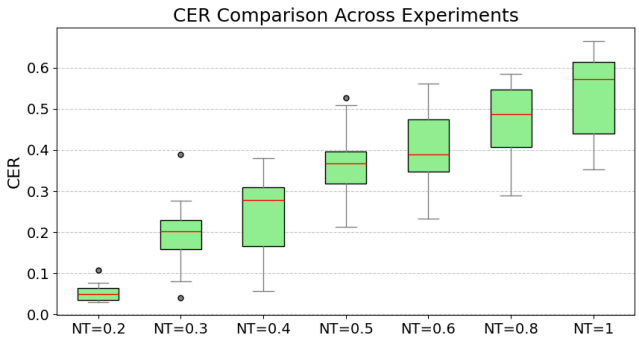


Figure 4: CER values for different noise thresholds. If the noise threshold is low, it cannot affect the audio enough. If the noise threshold is high, the attack becomes more successful while the noise becomes louder compared to the audio.

Experiments on the number of training samples					
NR	avg CER	var CER	avg WER	CER ASR	WER ASR
1	0.272	0.024	0.395	0.836	0.927
2	0.284	0.012	0.416	0.657	0.820
5	0.347	0.005	0.489	0.721	0.852
10	0.365	0.009	0.505	0.735	0.860
20	0.250	0.026	0.365	0.661	0.809
50	0.293	0.043	0.404	0.856	0.926
100	0.420	0.041	0.450	0.888	0.944

Table 2: Experiments on the influence of the number of training samples (NR). Unchanged parameters: compression factor = 10**, initial noise epsilon** = 0.02**, noise length** = $4 \cdot 16,000$**, noise threshold** = 0.5**, mutation rate** = 0.2**, elite fraction** = 0.2**, population size** = 100**, generations** = 30**, dropout probability to create noise bursts** = 0.95**.**

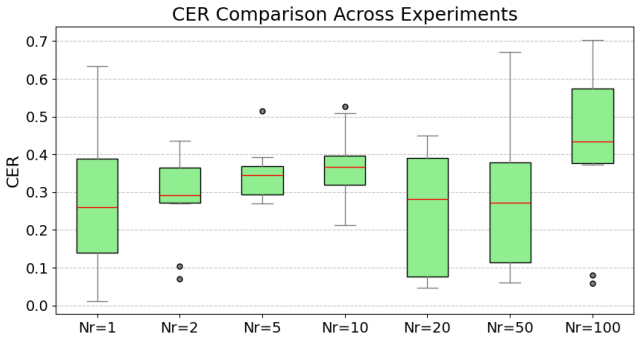


Figure 5: CER values for different numbers of training samples. Generalizability depends strongly on the number of samples used. Noise trained on just one sample already generalizes well. Still, generalizability otherwise increases with the number of training samples used, but the variance in the CER and WER values also increases.

because a higher threshold allows the noise to become louder and more bursts to be inserted since more noise is allowed in comparison to the original audio. For a noise threshold of 0.3, which allows the noise to have 30% of the original audio's energy, the attack is successful in terms of WER in over 70% of test cases. For a noise threshold of 0.5, the success is over 85%. This means that a balance needs to be found so that a human listener can still understand what is actually being said.

The goal of the universal attack is to find a single noise vector that causes mistranscriptions for any audio signal. The generalizability of perturbations makes the universal black-box attack much stronger than an individual attack. Although both require many queries during the training, the universal noise vector can afterwards be applied to any new input audio signal, while the individual noise vector only perturbs the known audio signal that it was trained on. To see how well a noise vector is able to generalize after training on a small subset of the data, the number of training samples is varied. Training on just 1 training sample is the individual attack from Section 4.3, but each individual perturbation vector is tested on all other test audio samples instead of just its training sample.

The results show that the generalizability of the noise vector to new audio signals is already possible for the individual perturbations of the individual attack that were trained on only one input audio signal, if the chosen training audio is suitable. This means that perturbations generated for a single audio sample are already surprisingly transferable and generalize well across all samples. A noise vector generated for a single training sample can perturb more than 20% of the words in over 90% of previously unknown audio samples. This suggests that adversarial weaknesses are shared between inputs in Wav2Vec, resulting in the transferability of individual perturbations to new audio signals even though no generalization is enforced by training on multiple audio inputs. In fact, the noise generated by the individual attack performs better than noise perturbations optimized in a small number of training samples, as can be seen in Figure 5 and Table 2. The previously presented individual attack is already a universal attack. A very large set of training data can produce even better results, as it captures more variety in the data, which generalizes better to new data in the test phase.

The attack is also tested on the newer state-of-the-art speech recognition model, Whisper, which is known to be very robust against noise in the audio and is computationally more expensive to train [9]. Here, the CER, WER, and attack success rates are much lower compared to Wav2Vec, but still show some attack success compared to clean Whisper transcriptions.

For Whisper, the attack does not receive as much meaningful information from the fitness values as for Wav2Vec. The fitness curves of Wav2Vec increase very fast in the beginning and converge early, while the fitness curves of Whisper show a slower, linear increase and converge only after many generations, which can be seen in the plots in Figure 6. Additionally, the fitness curves of Whisper can increase or decrease because Whisper is non-deterministic as a heuristic model. Thus, giving it the same input multiple times does not always result in the same output.

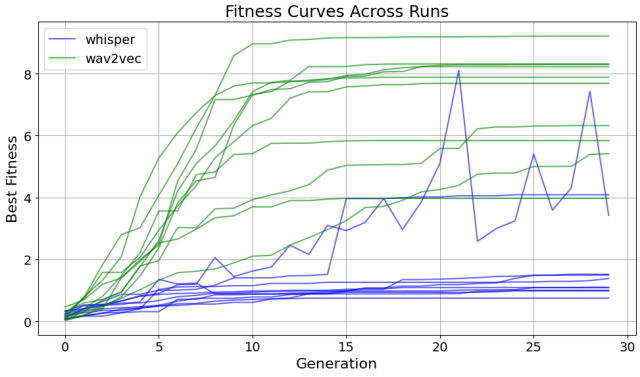


Figure 6: All fitnesses from the experiment on Whisper (blue lines) plotted against all fitnesses from the experiments on Wav2Vec (green lines). The fitness values from Whisper are consistently lower than the fitness values from training on Wav2Vec. The fitness curve of Wav2Vec grows on average much faster and then converges logistically, while the Whisper fitness curves grow linearly until they converge only after more generations.

When transferring the noise generated to attack Whisper to Wav2Vec, the attack success rate decreases significantly compared to the noise trained to attack Wav2Vec itself, as can be seen in Figure 7. However, the opposite, transferring the noise generated to attack Wav2Vec to Whisper, actually increases the average attack success rate on Whisper compared to its own noise. The maximum attack success rate is still achieved by Whisper-trained noise, but on average, Wav2Vec-trained noise outperforms its counterpart on Whisper, as shown in Figure 8.

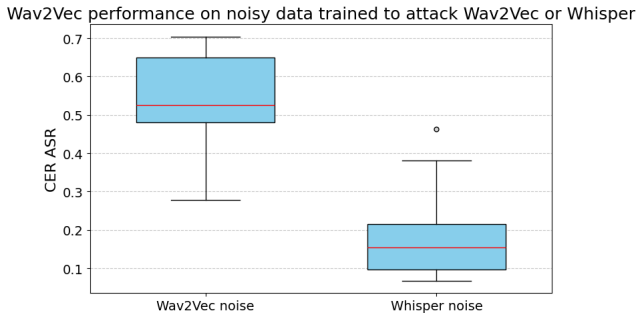


Figure 7: CER performance of Wav2Vec model on noise that was trained to attack Wav2Vec or Whisper. The noise trained to attack Wav2Vec causes more errors for Wav2Vec transcriptions than the one trained on Whisper.

This means that a transfer of noise between models is not only possible, but can be very successful if the noise is transferred from a weaker model to a more robust model. The weaker model shows very meaningful reactions to even small changes in the level of noise and is therefore very informative in its fitness function, while the robust model does not give as much meaningful feedback to the fitness function because the model does not react much to noise.

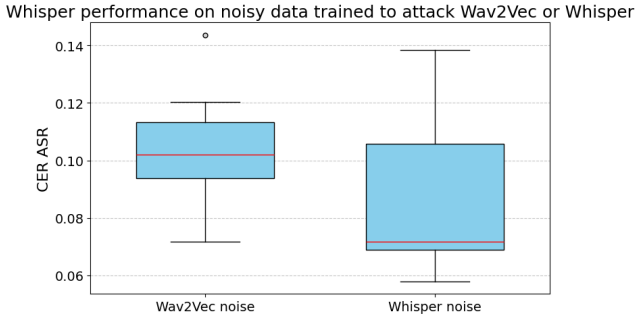


Figure 8: CER performance of the Whisper model on noise that was trained to attack Wav2Vec or Whisper. The noise trained to attack Wav2Vec is more consistent in causing errors for Whisper transcriptions than the one trained on Whisper (as the average is higher), but the highest error rates are achieved for noise trained on Whisper (as the maximum is higher).

The custom synthetic SMCP dataset has a significantly higher similarity score in terms of both part-of-speech tagging and Mel spectrogram L2 distance than the standard English Librispeech dataset. For a universal attack, which is designed to successfully attack multiple audio signals with the same noise vector, this property has major implications: since the data are more similar, the universal perturbations are more powerful because the data on which they are trained are more similar to the test data and therefore generalize better.

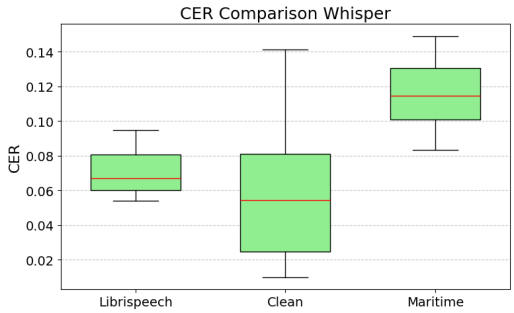


Figure 9: Character error rate of attack on maritime dataset applied on Whisper. Previous attack on Librispeech, clean transcriptions, and a new attack on synthetic SMCP data. Compared to Librispeech and the clean attack success rate on the maritime dataset, the new attack on the maritime dataset shows significantly higher success and much better generalizability of the universal perturbations.

Figure 9 shows that attacking the custom synthetic SMCP dataset with its higher similarity between audio files is significantly more successful than attacking the Librispeech dataset. When attacking Wav2Vec with this dataset, the increase is even more significant, as can be seen in Figure 10. This verifies that it is possible to achieve a consistently higher attack success rate on the more similar maritime

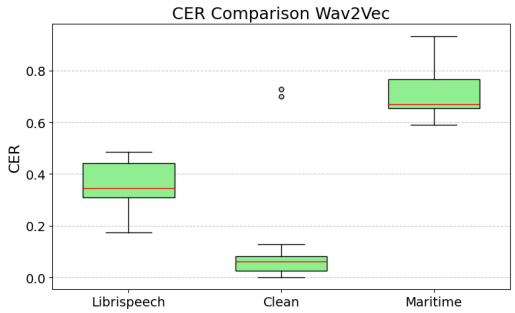


Figure 10: Character error rate of attack on maritime dataset applied on Wav2Vec. Previous attack on Librispeech, clean transcriptions, and a new attack on synthetic SMCP data. Compared to Librispeech and the clean attack success rate on the maritime dataset, the new attack on the maritime dataset shows significantly higher success and much better generalizability of the universal perturbations.

dataset, and that the universal attack generalizes better if the data is more homogeneous.

It is necessary to demonstrate that, since the attack produces audible noise, it does not disrupt a human listener's ability to understand what the original audio says. Therefore, a subjective listening test was performed with 19 participants. The details of ethical considerations and the instructions for participants are further discussed in Appendix B. The participants are presented with 8 audio samples, each of which is significantly mistranscribed by Wav2Vec. They are asked to rate how well they can understand the audio, how much effort it takes to understand the audio, and they transcribe the audio. Two of the audios use noise from experiments with the noise threshold set at 0.3, three use a noise threshold of 0.5, and three use a threshold of 0.8.

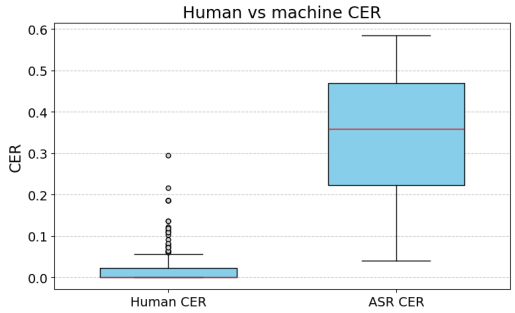


Figure 11: Character error rate of human transcripts versus ASR transcripts. ASR transcripts have a significantly higher error rate.

Compared to machine transcripts, the human listener makes significantly fewer errors when transcribing the audio, as can be seen in Figure 11. While the machine transcriptions have very high error rates and subsequently also high attack success rates, the human errors are far fewer, indicating that the human listener is still able to hear the speech in the audio, even though the machine is

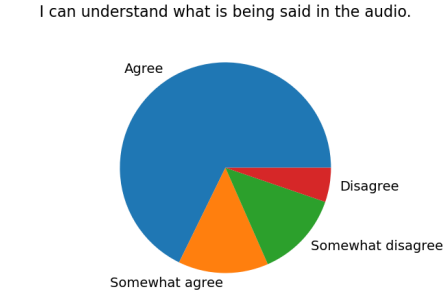


Figure 12: Pie chart showing ratio of agreement to the statement "I can understand what is being said in the audio."

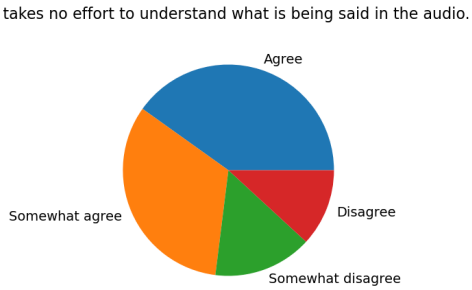


Figure 13: Pie chart showing ratio of agreement to the statement "It takes no effort to understand what is being said in the audio."

not able to do so anymore. This significant difference in error rates can also be observed when separating between noise thresholds, as for each one, the human-transcribed counterpart performs better, but humans have slightly more difficulty transcribing the noisier audio. Figures 12 and 13 show the ratio of each response about understanding and listening effort over all audio samples. Most participants indicate that they understand the audio without a problem and that it takes little to no effort to do so. The lower the noise threshold is set, the easier it is to understand what is said in an audio sample and the less effort is required for that. This indicates that while the noise is audible, it is not disruptive to a human's ability to understand the speech, even though it strongly disrupts the machine's ability to do so.

4.5 Semantic Defense

To distinguish between transcripts from clean and adversarially perturbed audio signals, a binary classifier was trained to predict the category from which a transcript originates. After 5 epochs, the model achieves consistently high performance, with an accuracy of 85.75%, precision of 93.08%, recall of 77.25%, and an F1 score of 84.43%. These results indicate that adversarial noise introduces noticeable linguistic artifacts in the transcripts, which the classifier

is able to distinguish only based on text data, without knowing the underlying audio data.

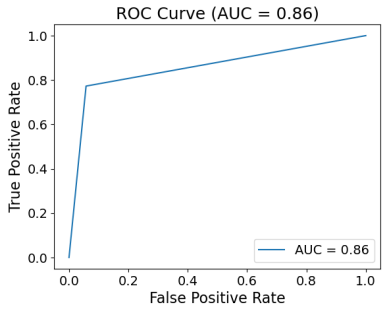


Figure 14: ROC curve for performance of linguistic defense on test data.

The ROC curve in Figure 14 shows that the binary classifier performs overall very well on the test data. The majority of test data points are classified correctly as either clean or adversarially perturbed transcripts. Still, there are noticeably more noisy examples that are falsely classified as clean than the other way around. The most likely explanation for this lies in the attack success rate: while the attack is highly successful on average, it does not cause mistranscriptions in all audios to which the adversarial perturbation is applied. Thus, it is likely that some of the training and test transcripts show little to no differences from clean transcripts, as the attack was not as successful on them, which is also why the classifier cannot distinguish them correctly. While this defense makes the assumption of a zero-knowledge scenario, where the attacker has no knowledge of the defense, we can still assume that the defense can work even against an adaptive attacker with knowledge of the text-based defense. The defense is a distinguisher of standard English text versus manipulated text that contains abnormal patterns. Since the attack is universal and untargeted, it is impossible to predict precisely how much of the original transcript of a previously unseen audio signal will be changed when the attack is applied and in which way. The attacker has no control over the changes; a well-trained adversarial perturbation can cause any changes. Since the changes are undirected, they are unlikely to produce real words, which is what the defense exploits. As long as the attack is untargeted, it is unlikely that an adaptive attacker could successfully change it to circumvent this defense. However, creating a targeted black-box universal attack with this same strategy while ensuring that human listeners can still understand what is being said is an extremely complex task, which makes the attack hard to adapt.

5 Discussion

The presented attack has been implemented successfully, showing the feasibility of the universal black-box adversarial attack on audio data using genetic algorithms that is deployed without any knowledge of the model internals and has thereby significant practical application in real-world commercial ASR systems, which are often not open-source. However, the demonstrated attack is not inaudible; if a listener knows what the attack sounds like, it is reasonably easy to hear it as background noise to the original clean audio. If the audio is already noisy in itself, which is typical for maritime radio, then the attack could become much more stealthy since the background noise is already expected. This means that under realistic conditions, the noise is less perceptible. There is a trade-off between attack success and audio quality. Finding the balance compared to the expected and given background noise level of the targeted audio signals is integral to real-world implementations of the attack, and consequently also to defenses against the attack.

Future research should try to make the attack less audible, for example by training a noise function instead of a fixed noise vector that reacts to streaming audio dynamically in real-time; this would be complex to train and to use in real-time, but has a lot of potential to be less audible. Another possibility could be applying a multi-objective optimization like in the individual black-box attack by Khare et al. [13] that returns a Pareto-front of text dissimilarity and audio similarity, to provide balanced candidate solutions.

The demonstrated attack uses phonetic targeting: the noise is generated specifically to mimic the short and sharp sound of obstruent phonemes. Future work in this area should include this feature explicitly by not just initializing random noise that suggests that the algorithm should evolve the random bursts that could be interpreted in this direction if the model evolves them in the right way, but by generating noise from a phoneme-matching function.

Comparing the models Wav2Vec 2.0 and OpenAI's Whisper, it becomes clear that Whisper is more robust to noise. On clean data, its character and word error rates are lower than for Wav2Vec under the same conditions, although by only a small margin, since Wav2Vec already performs well on the Librispeech dataset.

When transferring noise optimized on Wav2Vec to Whisper, the attack is on average significantly more successful. Transferring Whisper-optimized noise to Wav2Vec works, but is less successful than using its own Wav2Vec-optimized noise. This is most likely because Whisper is non-deterministic and does not give meaningful feedback to the attack optimization process. If the algorithm gets barely any feedback which noise features improve an attack and which do not, then it is not guided to improve the desired attack features in the right direction; it can only randomly try out combinations of which most give the same result, which is no attack success at all. When attacking Wav2Vec, the algorithm gets immediate feedback because the model is not as robust to noise, and it can therefore adjust the generated noise with much more precision. This is clearly visible in the fitness functions of the runs of the genetic algorithm: Figure 6 shows how for all experimental runs, Wav2Vec's fitness improves immediately from the start and grows rapidly within the first few generations until it converges early on. Wav2Vec's fitness provides meaningful feedback, so the improvements can be made with precision from the beginning. Whisper's fitness, on the other hand, improves on average in an almost linear line for many generations until it finds convergence much later than Wav2Vec, whereas the fitness of the attack on Wav2Vec is more similar to a logistic function. The absolute fitness values for the attack on Wav2Vec are also higher than for the attack on Whisper.

It is important to note that while Wav2Vec is deterministic and giving it the same audio signal multiple times will always produce the same transcript, Whisper, as a heuristic model, is not. This is clearly visible in some experiment runs for which the fitness function goes up and then down in the next generation, because

applying the same previously successful noise vector is not successful anymore when repeating the experiment. Therefore, Whisper does not only provide less meaningful feedback, but the feedback is also not consistent. A noise vector that shows a high attack success rate once might not give any attack success when used again, or vice versa. This behavior strongly influences the attack's ability to improve the fitness and thereby find a good universal perturbation. Because of this, the black-box optimization against a model as robust as Whisper is flawed by design. However, since the transfer of Wav2Vec-optimized noise to Whisper works well, this disadvantage can be circumvented by transferring from less robust deterministic models to more robust heuristic ones.

Given the early convergence of the feedback-rich Wav2Vec compared to the delayed convergence of Whisper, the fitness could potentially be used as a diagnostic tool to determine the vulnerability of a model against the attack. Future research should look further into proxy models to optimize a universal black-box attack instead of the more complex robust model. Using a simpler and deterministic proxy model is less computationally expensive and gives more meaningful feedback to optimize on. A challenge would be to find a proxy model that fits the target model well, but the results of the transferability experiments suggest that this method can improve attack success significantly compared to optimization on the robust target model itself. An interesting finding is that for this specific universal black-box attack, the individual attack trained on just one input is already successful as a universal attack. A noise vector trained on just one input already generalizes well to other previously unseen inputs. The experiment on the synthetic maritime dataset, which contains audio signals that are significantly more similar to each other than the audio data of the Librispeech dataset, confirms that the universal attack generalizes better across semantically and syntactically similar inputs, which is more likely the case for the formulaic structure of maritime communication, because the attack success rates are much higher for the attack on the maritime dataset. The smaller the phonetic variance is between training and testing samples, the more generalizable are the adversarial perturbations.

Future research on this could further analyze the speaker, the accent, and the vocabulary domain, for example, of nautical compared to general speech. It is also necessary to experiment on languages and multilingual models for application in the maritime domain which is known for communication in standardized nautical English but only if the conversation partners do not share the same native language; if they do, the conversation is more likely to be a mix of standardized English for specific vocabulary but the shared language for other parts of the conversation [11]. This rapid switch between languages is very difficult to follow for most ASR models.

The defense has been shown to be successful, but it classifies a fifth of the adversarially perturbed samples as clean. This could likely be because while the universal attack is a success in on average over 80% of cases, it is not always able to cause mistranscriptions. This means that some of the samples that are part of the noise class in both training and testing are indistinguishable from the clean class because no actual errors are present if the attack was not successful. This influences the defense's ability to correctly distinguish the two classes, of course. A way to improve the defense might be to exclude all noise class training data for which the attack was not successful. Then, the decision boundary can become more precise since no training data is ambiguous in its class, making the defense stronger despite not accurately representing all possible transcripts from the noise condition anymore.

The presented universal black-box attack has limitations. The attack is not imperceptible, but well audible as background noise, especially the more successful it becomes. Experiments have shown the importance of balancing attack success with the human listener's ability to still understand what is being said. The survey for human listeners has shown that a lower noise threshold is less disruptive and takes less effort to understand, but an increasing noise threshold is still understandable up to a certain point. The attack does not scale well to very robust models, as shown in the experiments on Whisper. However, adversarial perturbations transfer well from a less robust model like Wav2Vec to a more robust model like Whisper. The universal perturbations lose effectiveness if the domain and phonetic similarity drift further away from the training samples. This has been demonstrated by showing that the more similar synthetic maritime dataset achieves higher attack success rates. The measures of character and word error rate might not be meaningful in all cases. Librispeech consists of audio data taken from audiobooks, which means that it necessarily contains names for which there exists no normalized spelling. The reference transcripts from the dataset do not necessarily normalize American and British spelling of words like *"inquire"* and *"enquire"*, contributing to character and word error rates. Similarly, the synthetic maritime dataset contains names of vessels and coordinates, which also have ambiguous spellings. Short sentences also get immediately higher error rates for a smaller absolute number of errors than longer sentences, as the length of the audio files is not normalized. Compared to the human character and word error rates, it also does not take into account that human listeners intuitively fill gaps based on context, even though that might not be exactly what they hear. Just like visual illusions where humans see more than is present and fill in gaps, auditory illusions can skew how much a human actually hears and what they infer from the audio [20]. Because machines struggle to replicate this skill, comparing human and ASR transcriptions is difficult. Whisper partly addresses this by enforcing a subword vocabulary, but the uniquely human ability still skews results.

6 Conclusions

The presented universal black-box attack is highly relevant because communication-critical fields like the maritime domain are at risk from adversarial attacks that can be deployed in real-time. This attack has been shown to be successful for varying noise limits, training set sizes, transcription models, and in particular for more homogeneous datasets. We show that the individual attack trained on just one input is, in fact, already successful as a universal attack. We also demonstrate that Whisper is a more robust model against the attack. Finally, we show that more homogeneous data generalizes better from training to test data, making the universal perturbation more successful if it was trained on training data that is more similar to test data. A text-based defense was presented that can distinguish well between manipulated and unmanipulated transcripts.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.

[2] Amisha Rajnikant Bhanushali, Hyunjun Mun, and Joobeom Yun. 2024. Adversarial attacks on automatic speech recognition (asr): a survey. *IEEE Access*.

[3] Shobha Bhatt, Shweta Bansal, Ankit Kumar, Saroj Kumar Pandey, Manoj Kumar Ojha, Kamred Udham Singh, Sanjay Chakraborty, Teekam Singh, and Chetan Swarup. 2023. A comprehensive examination of phoneme recognition in automatic speech recognition systems. *Traitement du Signal*, 40, 5.

[4] Tanja Brcko Satler and Violeta Jurkovič. 2024. Solving maritime communication challenges with digimar: a practical approach. In *Maritime Transport Conference* number 10. Universitat Politècnica de Catalunya. Iniciativa Digital Politècnica.

[5] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*. IEEE, 1–7.

[6] Gianni D'Angelo and Francesco Palmieri. 2021. Gga: a modified genetic algorithm with gradient-based local search for solving constrained optimization problems. *Information Sciences*, 547, 136–162.

[7] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. 2020. Sirenattack: generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia conference on computer and communications security*, 357–369.

[8] Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du. 2023. Advddos: zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18, 3647–3661.

[9] Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4, 2.

[10] Hanqing Guo, Guangjing Wang, Yuanda Wang, Bocheng Chen, Qiben Yan, and Li Xiao. 2023. Phantomsound: black-box, query-efficient audio adversarial attack via split-second phoneme injection. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 366–380.

[11] Lamin Jatta. 2024. Maritime automatic speech recognition: improving the quality of transcriptions using artificial intelligence. *unpublished*.

[12] Eric Johansson. 2022. Using language models to improve a speech recognition based maritime emergency call detection system. *unpublished*.

[13] Shreya Khare, Rahul Aralikatte, and Senthil Mani. 2019. Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization. In *Proc. Interspeech 2019*, 3208–3212.

[14] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: a survey. *Information Fusion*, 102422.

[15] Christoph Martius, Emin Çağatay Nakilcioğlu, Maximilian Reimann, and Ole John. 2024. Refining maritime automatic speech recognition by leveraging synthetic speech. *Maritime Transport Research*, 7, 100114.

[16] Emin Nakilcioglu and CML Fraunhofer. 2023. Automatic speech recognition in the maritime domain. In *Autonomous Ship Expo and Conference 2023*.

[17] Emin Cagatay Nakilcioglu, Maximilian Reimann, and Ole John. 2023. Adaptation and optimization of automatic speech recognition (asr) for the maritime domain in the field of vhf communication. In *International Conference on Computer and IT Applications in the Maritime Industries 2023*.

[18] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. 2019. Universal adversarial perturbations for speech recognition systems. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2019, 481–485.

[19] Douglas O'Shaughnessy. 2024. Trends and developments in automatic speech recognition research. *Computer Speech & Language*, 83, 101538.

[20] Angélique A Scharine and Tomasz R Letowski. 2009. Auditory conflicts and illusions. *Helmet-mounted displays: sensation, perception and cognition issues*, 579–598.

[21] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2019. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society.

[22] Zheng Sun, Jinxiao Zhao, Feng Guo, Yuxuan Chen, and Lei Ju. 2024. Commanderuap: a practical and transferable universal adversarial attacks on speech recognition models. *Cybersecurity*, 7, 1, 38.

[23] Hao Tan, Le Wang, Huan Zhang, Junjian Zhang, Muhammad Shafiq, and Zhaoquan Gu. 2022. Adversarial attack and defense strategies of speaker recognition systems: a survey. *Electronics*, 11, 14, 2183.

[24] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. 2019. Targeted adversarial examples for black box audio systems. In *2019 IEEE security and privacy workshops (SPW)*. IEEE, 15–20.

[25] Jiakai Wang, Zhendong Chen, Zixin Yin, Qinghong Yang, and Xianglong Liu. 2022. Phonemic adversarial attack against audio recognition in real world. *arXiv preprint arXiv:2211.10661*.

[26] Darrell Whitley. 2001. An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and software technology*, 43, 14, 817–831.

[27] Xiaoliang Wu and Ajitha Rajan. 2022. Catch me if you can: blackbox adversarial attacks on automatic speech recognition using frequency masking. In *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 169–178.

[28] Hiromu Yakura and Jun Sakuma. 2018. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793*.

[29] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. 2021. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*.

[30] Yujun Zhang, Yanqu Chen, Jiakai Wang, Jin Hu, Renshuai Tao, and Xianglong Liu. 2025. Generating targeted universal adversarial perturbation against automatic speech recognition via phoneme tailoring. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[31] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. 2021. Black-box adversarial attacks on commercial speech platforms with minimal information. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, 86–107.

A Datasets

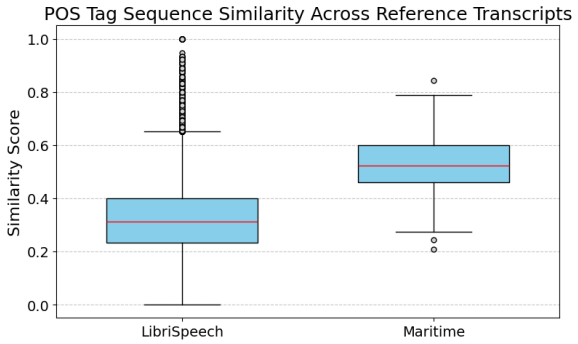


Figure 15: Part-of-speech tagging similarity between Librispeech and the synthetic maritime dataset. It can be seen that the synthetic SMCP data is significantly more homogeneous compared to the standard English Librispeech data.

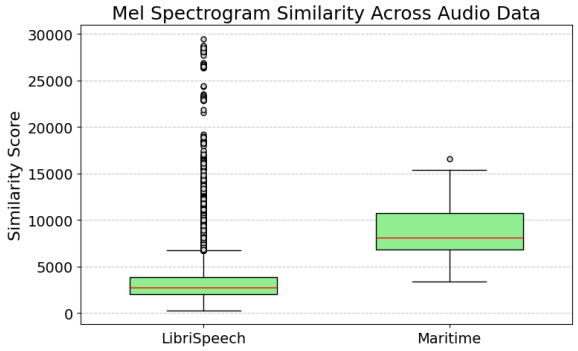


Figure 16: Mel spectrogram L2 similarity comparing Librispeech and the synthetic maritime dataset. It can be seen that the synthetic SMCP data is significantly more homogeneous compared to the standard English Librispeech data.

The Librispeech dataset has the advantage of containing a large amount of clean data of varying lengths without much background noise. However, for the application of this research, it is also interesting to demonstrate the attack on data that is derived from the specific SMCP protocol which contains many standardized phrases and expressions. Since there is no suitable SMCP dataset publicly available, a small dataset is instead synthesized by generating 30 examples of maritime radio communication phrases in emergency scenarios and creating audio files from them with Luvvoice's text-to-speech technology. Although this small dataset can only be seen as an approximation of real maritime radio communication, it incorporates the very specifically structured nature of SMCP phrases for emergencies and showcases that the typical sentences communicated via maritime radio are more similar to each other than standard English sentences, which can have an effect on the attack.

This increased similarity can be shown both in terms of part-of-speech (POS) tagging and mel spectrogram similarity between audio signals, as seen in Figures 15 and 16. The Mann-Whitney U test confirms that the syntactic structure of the sentences in the synthetic maritime dataset is significantly more homogeneous than in the Librispeech dataset ($U = 133,485,300.5$, $p = 2.31 \cdot 10^{-193}$). Similarly, the Mann-Whitney U test confirms that the audio signals in Mel spectrogram form are significantly more similar to each other in the SMCP dataset than in the Librispeech dataset ($U = 2,637,412$, $p = 1.69 \cdot 10^{-232}$). Thus, the similarity scores for both measures of similarity are significantly higher for the synthetic SMCP dataset, making the audio signals and spoken sentences in the SMCP dataset more similar to each other than in the standard English Librispeech dataset. Even though the dataset is small compared to the selected subset of Librispeech of 2620 samples, experiments can show if there is a difference in universal attack performance if the data is more homogeneous.

B Qualitative survey for human listeners

To show that the noise threshold for the adversarial perturbations has been chosen suitably, an online subjective listening test was conducted with 19 volunteers. The survey was completely anonymous. Each participant had to record their informed consent to the research and could opt out at any point. Before starting, the participants were extensively informed about anonymity, the purpose of the study, what kind of data will be gathered, how any data would be used, how to withdraw from the study, the benefits, discomforts, and risks of the study, contact details so they could get more information, and what they had to do during the survey. The invitation to participate was given to university students studying for a full-time degree taught in English that requires proof of language skills to get in, therefore a sufficient level of English to complete this survey is assumed.

The participants are presented with 8 audio samples, each of which is significantly mistranscribed by Wav2Vec. The estimated time for the survey was 5 minutes. For each of the 8 audio samples, the participants are first asked to rate two statements on a Likert scale and then to transcribe the audio themselves. To do so, they can hear the audio at least 3 times (once for each question), and potentially more often, to write down what is said in the audio. The first statement is: "I can understand what is being said in the audio." The second statement is: "It takes no effort to understand what is being said in the audio." The Likert scale offers four points of reference: "Agree", "somewhat agree", "somewhat disagree", and "disagree". Two of the audios use noise from experiments with the noise threshold set at 0.3, three use a noise threshold of 0.5, and three use a threshold of 0.8.

The survey is not associated with any risks or discomforts greater than what is encountered in daily life while listening to audio data, which all participants were informed about. No benefits were associated with participation. No personal data about the participants is gathered apart from their consent to the research and their confirmation that they are above 18 years old to participate. Every other question only asks about the participant's ability to listen to the audio. Participation is completely anonymous. The purpose of the study was explained to the participants, and they were given contact details to ask for more information if desired. They were informed how their data would be used, stored, and processed.