# Towards Robust Muting Attacks on Whisper

*Lucas Tucker*[1], *Jose Alec Tandoc*[1]

[1]University of Chicago

{lwtucker, atandoc} @uchicago.edu

## Abstract

Adversarial examples are inputs to machine learning models designed to produce incorrect outputs. In automated speech recognition (ASR), this involves crafting audio samples whose machine transcriptions have high word error rate relative to respective human transcriptions. This paper builds upon a specific type of attack proposed by Raina and Gale [1] wherein the attacker prepends a small $\sim$1 second clip of audio to any source audio file and OpenAI's Whisper ASR model transcribes the resulting combined audio as an empty string. In particular, we find that prepending gaussian noise to the attacked audio recovers the transcription, and we design a more flexible universal attack that may be placed anywhere within the first $30,000$ frames of audio to mute OpenAI's Base Whisper ASR model on audio clips under 20 seconds.

## 1. Introduction

An adversarial example for a machine learning model is an input $x'$ that bears resemblance by some metric to a "benign" sample $x$ yet is classified differently by the machine learning model. In the audio recognition space, adversarial examples are audio samples $x'$ whose human recognizable transcription differs from that of machine transcription by some word error rate (WER) threshold, which is $30\%$ in our case. These samples $x'$ involve some added noise or perturbation from the original sample $x$, although the placement and extent of this noise differs largely between attacks [2] [3] [4]. By "white-box" attacks, we refer to the construction of such adversarial examples that make use of the model weights and architecture. Targeted adversarial examples in automatic speech recognition are those modified $x'$ inputs that transcribe to a phrase of the attacker's choice [5], and by "universal" attacks we refer to audio modifications that produce adversarial examples regardless of the source audio under a specific automatic speech recognition model. A 2024 paper from Raina and Gale [1] proposed the creation of a universal white-box prepend attack to "mute" OpenAI's Whisper transcription model, wherein the attacker prepends a 1 second audio clip to any source audio and Whisper transcribes the combined audio as an empty string. In this paper we ask how to defend against such an attack and how to design a more robust attack. We show that:

- Prepending under 500 frames of noise to Raina and Gale's "muting" adversarial audio recovers transcription.
- Varying embedding placement during training of a "muting" adversarial segment increases its robustness to prepended noise on under 20 second audio clips.
- The length of the source audio being attacked is highly relevant to the success of the attack due to Whisper's segmentation of mel spectrogram tokens.

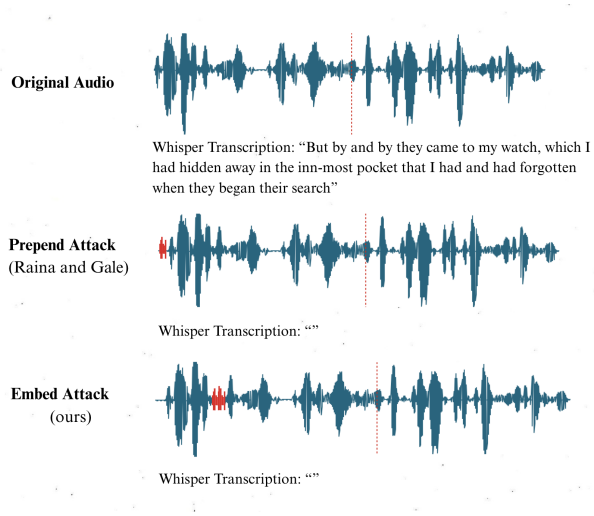


Figure 1: *Visual representation of attacks*

## 2. Background

### 2.1. Whisper's Autoregressive Decoding Structure

OpenAI's multitask recognition model Whisper [6] utilizes special tokens to determine whether to perform transcription or translation, what language to use, and where in the audio the transcription is taking place. Whisper segments audio into up to 30-second segments, corresponding to 3000 frames of the mel spectrogram, decoding each with small overlap. Decoding is performed autoregressively, where the likelihood of a token sequence $\mathbf{y}$ given task $T_{\text{transcribe}}$ and segment $\mathbf{x}$ is as follows:

$$P_\theta(\mathbf{y} \mid \mathbf{x}, T_{\text{transcribe}}) = \prod_{i=1}^{m} P_\theta(y_i \mid \mathbf{y}_{<i}, \mathbf{x}, T_{\text{transcribe}}) \quad (1)$$

where $\theta$ denotes model parameters. In particular, at each step $i \in [m]$ the model outputs a distribution over 51865 tokens, the maximum of which is transcribed. Whisper concludes its transcription with an "end of text" (EOT) token. Critically, the "segment" $\mathbf{x}$ is a segment of the encoded mel-spectrogram of the source audio. Whisper tracks where the current segment lies using a "seek" variable.

### 2.2. Projected Gradient Descent (PGD) for Audio

Projected Gradient Descent is a white-box optimization method used to design adversarial examples by iteratively updating an initial bit of noise using the model gradient. In our case, we wish to maximize the probability that Whisper first transcribes an "end of text" token. Let $f : \mathcal{X} \to \mathcal{Y}$ denote the Whisper model as a function mapping from audio space to textual/transcription space, and let $g : \mathcal{X}^2 \to \mathcal{X}$ denote the em-

bedding of the adversarial audio into a source segment. Then, updates are performed to the adversarial segment $\mathbf{x}^{(i)}$ as

$$\mathbf{s}^{(i+1)} = \mathbf{x}^{(i)} - \nabla_{\mathbf{x}^{(i)}} \mathcal{L}(f(g(\mathbf{x}^{\text{source}}, \mathbf{x}^{(i)})), \text{“”}) \quad (2)$$

$$\mathbf{x}^{(i+1)} = \text{CLAMP}(\mathbf{s}^{(i+1)}, -\epsilon, \epsilon) \quad (3)$$

for some fixed $\epsilon$ value corresponding to decibels, and where “” corresponds to the transcription that begins with the terminating EOT token mentioned above. We use $\oplus$ to refer to concatenation of audio, so that in equation (2) we split the source audio at its $n$-th frame and insert $\mathbf{x}^{(i)}$ in between.

## 3. Methodology

Building off of the work of Raina and Gale [1], we found that their universal prepend attack relies on decoder attention toward the start of the encoded first audio segment, and that prepending a few hundred frames of Gaussian noise to the attack segment recovered the transcription. If we let $\mathbf{w}$ denote the prepend segment and $\mathbf{g}$ denote 200 frames of Gaussian noise, we observe that while

$$\mathbf{x}^{\text{adversarial}} = \mathbf{w} \oplus \mathbf{x}^{\text{source}}$$

yields empty transcription,

$$\mathbf{x}^{\text{restored}} = \mathbf{g} \oplus \mathbf{w} \oplus \mathbf{x}^{\text{source}}$$

can yield the original transcription. In particular, prepending such Gaussian noise reduces the attack success rate by nearly $80\%$, where we have used the pre-computed prepend attack segment provided by the source code of Raina and Gale. Our goal was to create a new universal segment $\mathbf{x}^*$ that is robust to prepended noise.

To remove a dependence on the time-position of the attack, we trained an adversarial audio clip inserted variously within the first $30,000$ frames of source audio samples to increase the value that the first token predicted after this token prefix is the “end of text” token. Note that Whisper begins decoding with a token prefix consisting of the “start of transcript” token, language tag, task tag, and “no time stamps” tag. Formally, let $\mathbf{y}_0$ denote this token prefix, denote the $N$ training audio samples $\mathbf{x}^j$ of LibriSpeech, and let $\ell$ denote the fixed *segment* size we choose for training. Then, for the embedding cutoff $n_j \sim 1000 \cdot \text{Unif}(\{1, 2, ..., 30\})$ we define the $j$-th embedded clip as

$$g(\mathbf{x}^j, \mathbf{x}) = \mathbf{x}^j_{0:n_j} \oplus \mathbf{x} \oplus \mathbf{x}^j_{n_j:\ell} \quad (4)$$

using the same notation as in equation (2). We performed training based on the objective function

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmax}} \prod_{j=1}^{N} P_\theta(y_1 = t \,|\, g(\mathbf{x}^j, \mathbf{x}), \mathbf{y}_0) \quad (5)$$

where

$$t = \langle|\text{endoftext}|\rangle$$

Importantly, in our definition of (4) we sample cutoffs $n_j$ randomly to ensure robustness to placement of the adversarial noise anywhere within the first $30,000$ frames of audio. A key assumption in this training is the fixed segment length $\ell$, since $\ell = 3000$ if the audio sample is at least 30 seconds, but we opted to fix $\ell$ at a lesser value to accomodate the smaller audio clips we used in LibriSpeech. We observed worse performance when we did not fix $\ell$ and instead allowed varied audio length input to the decoder. We trained using PGD and found that clipping with $\epsilon$ values decreased performance, so we ultimately opted to let $\epsilon$ be a large value while keeping the adversarial segment $\sim$0.64 seconds in length. We performed one gradient step as in equation (2) per audio clip over 5 epochs, which took 30 minutes to run on an RTX 4070 GPU.

## 4. Results

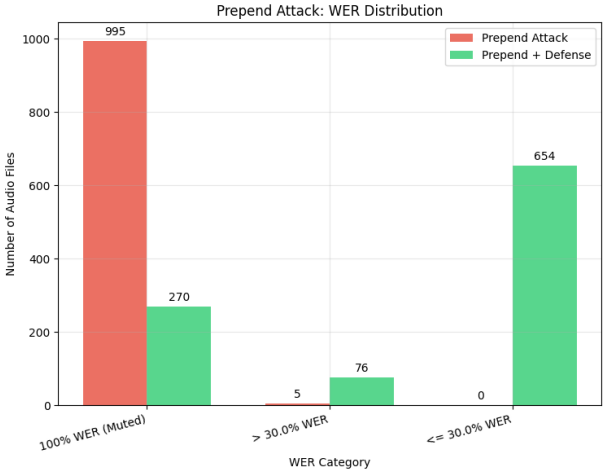


Figure 2: *Relative WER (Compared to Whisper's transcription on Benign): Prepend attack*

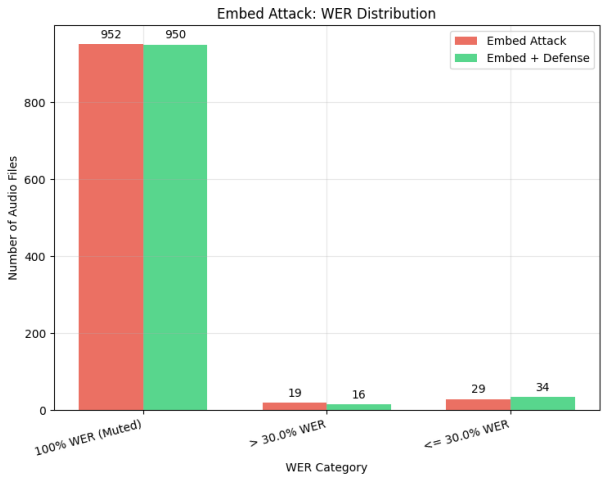


Figure 3: *Relative WER (Compared to Whisper's transcription on Benign): Embed attack*

### 4.1. Attacks on Benign and Defended Audios

We demonstrate the effectiveness of our attack compared to the prepend attack and against the Gaussian noise defense. We split the results into three sections: 100% word error rate (WER), $\geq$ 30% WER, and $<$ 30% WER. Here, we are comparing relative word error rate to Whisper's transcription of the benign audio, where 0% implies that the attacked audio was exactly transcribed as the benign audio. 100% WER implies

that a transcription was fully muted, while chose 30% WER as a threshold of intelligibility. If a transcription is not muted, but has a high enough WER, the adversarial insertion is still considered an effective attack, as it damages the transcription enough to ruin the intelligibility of the script as compared to the Whisper's original transcription.

As shown in Figures 2 and 3, both the prepend and embed attacks saw over 95% fully muting transcriptions. However, once our smaller 50-frame Gaussian prepend defense is applied, the prepend attack only worked less than 30% of the time in fully muting the transcriptions. However, that same defense applied to our embed attack saw little change in muting success, suggesting that prepended noise maintains or better mimics our objective function. By "Prepend + Defense" we refer to prepend attacks with $k$ frames of Gaussian noise inserted in front, where in Figures 2 and 3 we fix $k = 50$ while in Figures 4 and 5 we sampled $k \sim \text{Unif}([100, 500])$

| Method | Average WER |
|---|---|
| Prepend | 99.89% |
| Prepend + Defense | 38.07% |
| Embed | 96.74% |
| Embed + Defense | 96.61% |

Table 1: *Comparison of Methods by Relative WER*

As can be seen in the above table, the average WER on the prepend attack drops significantly, indicating that we are indeed recovering transcripts effectively with the defense. The embed attack is far more robust against the defense, with average WER changing by less than 0.20%.

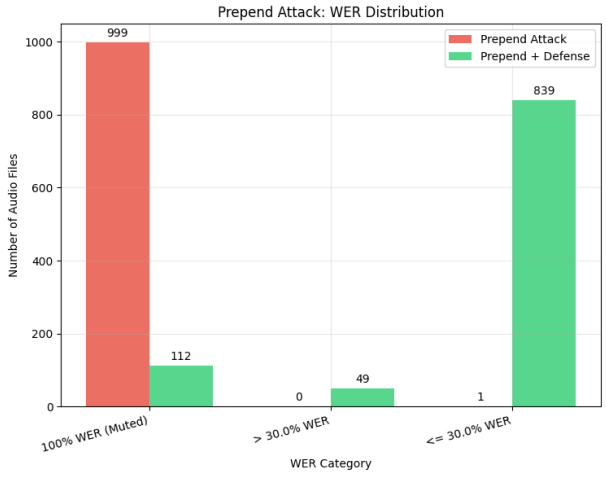


Figure 4: *Prepend attack: Gaussian Defense varying from 100-500 frames*

### 4.2. Varied Gaussian Defense

As mentioned, we further examined the robustness of these attacks against a new attack, with Gaussian noise of $100 - 500$ varying frame size. We can see that the higher noise volume was more effective against the prepend attack. Interestingly, this higher noise volume was less effective against our embed attack, as we witness an increase in fully muted transcriptions compared to a no-defense scenario.

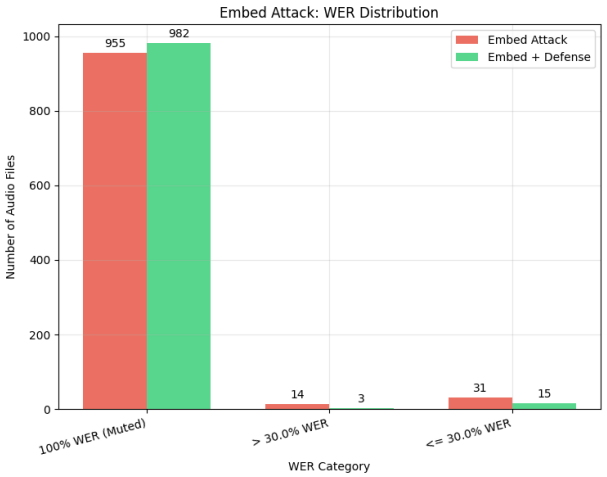


Figure 5: *Embed attack: Gaussian Defense varying from 100-500 frames*

| Method | Average WER |
|---|---|
| Prepend | 99.92% |
| Prepend + Defense | 19.28% |
| Embed | 96.71% |
| Embed + Defense | 98.55% |

Table 2: *Comparison of Methods by Relative WER with varying 100-500 frame Gaussian Noise Defense*

## 5. Discussion

### 5.1. Embed Attack Robustness to Gaussian Noise

We can see that our embed attack is similarly effective to the prepend attack in the benign scenario, yet robust in the defense scenario. We hypothesized that this is because we have trained an attack that is robust to varying noise within the first segment. In both the prepend and embedding scenarios, these attacks rely heavily on the autoregressive nature of Whisper. As previously mentioned, Raina and Gale's objective function optimizes for the 0th frame to be predicted as EOT. Because this objective is strictly optimizing for the beginning of the sequence, any time shift, even as small as 50 frames, creates an out-of-domain problem for the prepend attack, as the objective function was not designed to handle the extra noise it is sees between the prepend sequence $\mathbf{y}_0^*$ and the adversarial attack. Due to the time-shift, the adversarial audio no longer forces the model to attend to the prepend sequence $\mathbf{y}_0^*$, so EOT is no longer predicted. We compare this to the objective function of our embed attack, equation (5). By ranging the cutoff we effectively introduce random noise between our prepend sequence and our adversarial attack. This additional noise pushes the adversarial audio into an EOT predicting minima agnostic to prior noise.

### 5.2. Initial Segment Placement

An important note that we have mentioned for both of these attacks is that they must occur in the first "segment" of au-

dio, where the segment length is dynamically determined by Whisper at inference, based on the length of the input audio. As mentioned in the background, Whisper's encoding-decoding method converts an input audio into a mel-spectrogram and then segments it into overlapping windows of size at most 30 seconds, which is then decoded segment by segment. For most audios, this segment length is typically smaller than 30 seconds, with our estimate being 128,000 frames, around 8 seconds. Thus, we threshold our cutoff much lower than this, at 30,000 frames, to ensure that we are within the first segment. Placement within the first segment is essential to our attack because of how we predict the EOT token. These adversarial audios are optimized to force the model to heavily attend to $\mathbf{y}_0^*$ and predict EOT. However, the prepend sequence $\mathbf{y}_0^*$ only exists within the first segment, so if our adversarial audio is placed outside of there, it will be impossible for the model to attend to $\mathbf{y}_0^*$.

### 5.3. Embed Performance Increase with Varied Gaussian Noise

We extended our defense to Gaussian noise varying between 100-500 frames, with the hypothesis that larger time shifts would make a better defense. We saw that this was was true for the prepend attack. Unsurprisingly, this did not worsen the efficacy of the embed attack, as we trained it to be robust to time shifts. However, what is interesting is that we see a slight improvement in results when the defense is applied, as compared to the undefended audio. This may be due to the training process of our embed attack, which is sampling extra audio that could come from a distribution similar to the Gaussian noise we are applying as a defense. However, this is not something we can confirm with the limited computational power and time, but is something to consider for future work.

## 6. Future Work

### 6.1. Segment Agnostic Embedding

Our current embedding relies on remaining within the first segment of the transcription audio. Exploring a placement-agnostic adversarial attack would be the next best step in this process.

### 6.2. Embedding Efficacy under varied Gaussian Noise Defenses

As mentioned in discussion, we are unsure of the effects of high volume Gaussian noise defenses on our embedding attack. Exploring the robustness of this attack towards our proposed Gaussian defense is important, as it may mean that a different defense will be needed to defend this attack.

## 7. Conclusions

We find that prepend-only attacks are susceptible to time-offset defenses, such as additional prepended Gaussian noise. However, we see that our embed attack has demonstrated a significant improvement against the original prepend defense. Both of these attacks demonstrate a key issue with modern ASR models. They are not robust towards adversarial attacks and are easily manipulated, spelling trouble for systems that may use ASR transcriptions for content filtering or other security concerns.

## 8. Acknowledgements

We would like to thank Professor Karen Livescu, Ju-Chieh Chou, Chung-Ming Chien, and the Toyota Technical Institute at Chicago for their support.

## 9. References

[1] V. Raina, R. Ma, C. McGhee, K. Knill, and M. Gales, "Muting whisper: A universal acoustic adversarial attack on speech foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2405.06134

[2] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," 2019. [Online]. Available: https://arxiv.org/abs/1903.10346

[3] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023. [Online]. Available: https://arxiv.org/abs/2307.15043

[4] V. Raina and M. Gales, "Controlling whisper: Universal acoustic adversarial attacks to control speech foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2407.04482

[5] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," *2018 IEEE Security and Privacy Workshops (SPW)* , pp. 1–7, 2018.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356