Dynamic Epsilon applied Masks in Unlearnable Audio

Abstract:

Within recent years, progress has been made to prevent machine learning models from training on data that is copyrighted, malicious, or contains personal data. This data becomes unlearnable through the introduction of error-minimizing noise. While previous research has mostly made progress regarding adding imperceptible noise to images and LLMs, there is significantly less progress in audio unlearnability. Our research involved adding error-minimizing noise to audio samples by splitting wavelengths into segments and applying a normalized mask to each segment based on its average amplitude- which represents a fraction of the epsilon. This allows for more noise in sections of stronger amplitude, and thus, increases unlearnability without adding perceptible noise during the quieter segments.

Setup and Datasets:

Our audio data was obtained from torchaudio's Speech Command Classification dataset. This contains 32 speech commands of 1 second duration each (in the case of our testing, only 13 of these speech commands were used). Our model trains on 30639 samples and tests on 3970 samples- classifying each sample into one of the 13 classes.

Segmented Wavelengths and Dynamic Epsilon Values:

The goal of our research was to make the model unlearnable by implementing error-minimizing noise into the training samples, while also maintaining imperceptibility to human hearing. Our method of preventing noticeable noise involved the use of a non-constant epsilon value. This enables the noise that is added to vary depending on the amplitude of the wavelength.

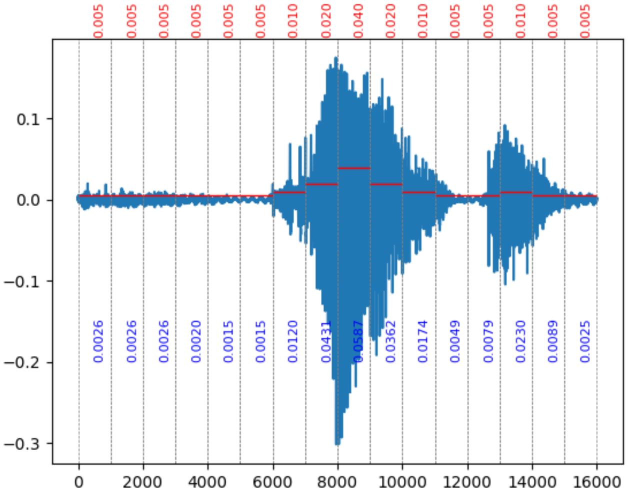


Figure 1: Segmented Wavelength

Each audio sample has a sampling rate of 16000 hz and is split into segments of a specified size. Within each segment, the average amplitude is calculated, and a mask is applied depending on the strength of the segment. This allows weaker, quieter sections of the sample to have less noise applied, while louder sections of the sample have increased noise. A constant, maximum epsilon is then multiplied by each segment's mask- normalized from 0 to 1.

As seen in Figure 1, each wavelength was split into segments of size 800 Hz. The average amplitude (blue value) is used to determine the mask value. These masks are then multiplied by the max epsilon and result in a fraction of this epsilon (red value).

Tests/Results:

CONSTANT EPSILON

To achieve unlearnability, an epsilon value of 0.08 is applied to the entire wavelength. This results in an accuracy of around ~14.156%, indicating the model is unable to learn much from the perturbed data. However, this level of noise is quite noticeable. The comparison between a clean wavelength and a perturbed one can be seen in Figure 2 below.

In the audio classification, quieter sections (that often contain no speaking) are less important in terms of classifying the word spoken. Therefore, the noise applied during these sections has little effect on the unlearnability of the audio. Contrarily, the noise in these sections has a large impact on the imperceptibility- both visually and audibly.

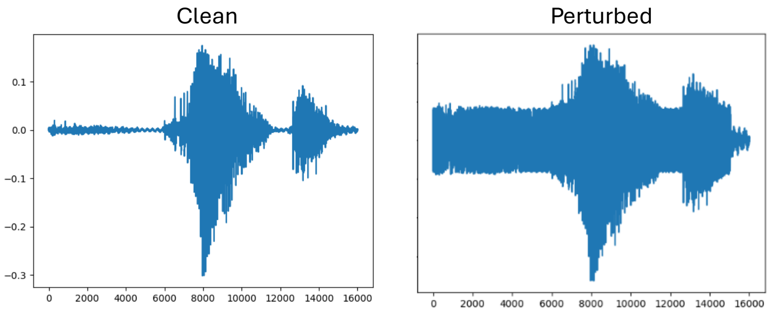


Figure 2: Constant 0.08 Epsilon

DYNAMIC (SEGMENTED MASK) EPSILON

To achieve both imperceptibility and unlearnability across a variety of amplitudes, we used varying epsilon values. To begin, the wavelength is split into segments (our testing involved segments of size 1000 Hz and 800 Hz, equaling 16 and 20 segments per wavelength, respectively). Figure 1 shows an example wavelength of segment size 1000 Hz. The average amplitude in each of these sections is calculated and used to determine the mask values for each segment. These masks- ranging from 0 to 1- are used to determine what fraction of the

Amplitude	Epsilon Mask	Epsilon (eps = 0.13)
> 0.1	1.0000	0.13
0.08 < A ≤ 0.1	0.5380	0.07
0.05 < A ≤ 0.08	0.3077	0.04
0.03 < A ≤ 0.05	0.1538	0.02
0.01 < A ≤ 0.03	0.0769	0.01
< 0.01	0.0385	0.005

Table 1: Segmented Masks

epsilon is applied. Table 1 contains the mask values applied across varying amplitudes, and the corresponding epsilon value when max epsilon is set to 0.13 (as used in our testing)

These mask values- which are unique to each audio sample and its segments- are then multiplied by a constant epsilon value that represents the maximum epsilon across the audio samples. Any segments with an average amplitude of above 0.1 will be given the maximum noise, while any segments under 0.1 will be given a fraction of this epsilon. Our Min-Min function updates the noise on each segment individually but combines these segments together when calculating the loss on the model.

ACCURACY RESULTS

Using the mask values from Table 1 with a max epsilon of 0.13, results in the following perturbed wavelength (Figure 3). When training on the perturbed audio below, our model has an accuracy of approximately 14.005% across 10 epochs- comparable to the accuracy of a constant 0.08 epsilon applied above. While still not completely unlearnable, it is significantly lower than the clean accuracy of approximately 85%, while being significantly less noticeable than the noise in the constant epsilon from Figure 2.

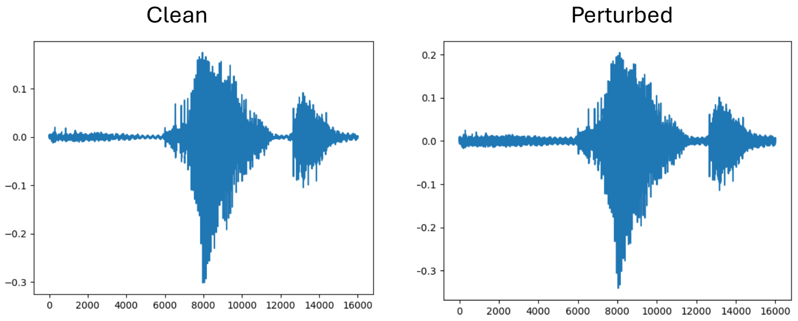


Figure 3: Dynamic Epsilon (eps = 0.13)

In addition, resampling from 16 kHz to 8 kHz (and back) is tested as well on the perturbed data. Under this transformation, it still achieves a low accuracy- at around 18.32%. The accuracies of these different tests are in Figure 4.

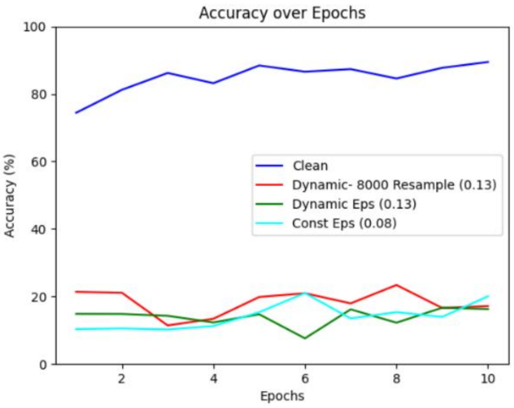


Figure 4- Accuracies

IMPERCEPTIBILITY RESULTS

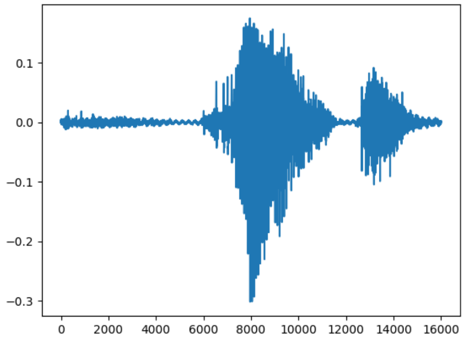
As shown above, using segmented epsilon values resulted in less noticeable noise while still maintaining a low accuracy of around 14%. Both the perturbated examples in Figure 2 and Figure 3 have approximately the same unlearnability, but the dynamic method of applying epsilons is significantly less perceptible.

Audibly, however, mid-range amplitudes (0.07-0.2) still have somewhat noticeable noise. Further tests and improvements could be made in this area. Though it still is significantly less noticeable than the previous method of constant epsilons. (The WAV files are attached below and on the GitHub)

DATA

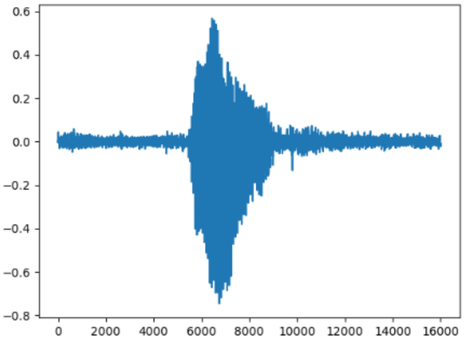
Clean Data (Accuracy = about 85%):

Sample 1 (Clean)



clean_0.wav

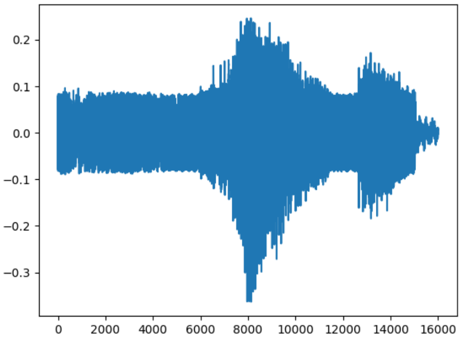
Sample 2 (Clean)



clean_1.wav

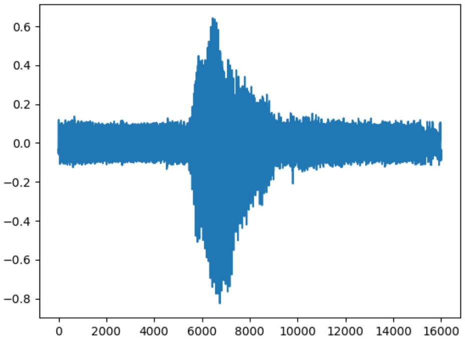
Constant Epsilon (Epsilon 0.08 applied to entire wavelength, Accuracy = about 14.156%):

Sample 1 (Const Eps = 0.08)



noise_0.wav

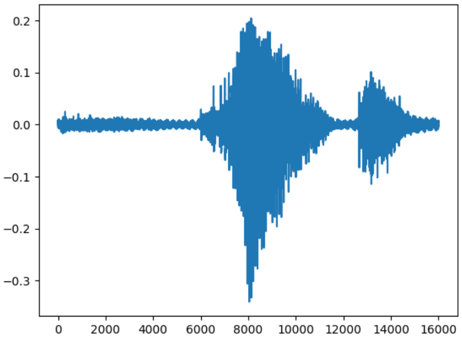
Sample 2 (Const Eps = 0.08)



noise_1.wav

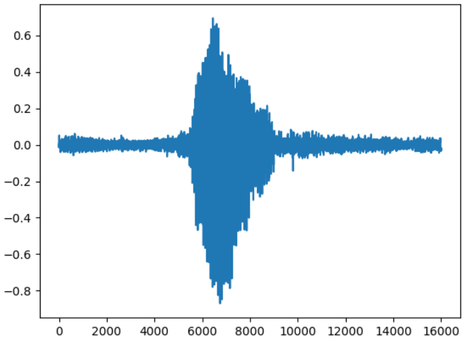
Dynamic Epsilon (Max Epsilon = 0.13, Accuracy = about 14.005%):

Sample 1 (Dynamic, Max Eps = 0.13)



noise_0.wav

Sample 2 (Dynamic, Max Eps = 0.13)



noise_1.wav

For further results and code used, see ReaganSanz/MoSIS-Unlearnable-Audio: Code and research from Summer 2024 REU Internship at MoSIS Lab in Audio Unlearnability (github.com)

References:

Huang, Hanxun, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. "Unlearnable Examples: Making Personal Data Unexploitable." *Proceedings of the International Conference on Learning Representations* (ICLR), 2021

PyTorch. “Speech Command Classification with Torchaudio.” Speech Command Classification with Torchaudio - PyTorch Tutorials 1.13.1+cu117 Documentation, 2022, pytorch.org/tutorials/intermediate/speech_command_classification_with_torchaudio_tutorial.html