Psychoacoustic Adversarial Attacks on Speech Recognition Systems

Tomer Erez, tomer.erez@campus.technion.ac.il

Department of Computer Science – Technion, Haifa, Israel
https://github.com/tomer-erez/Psychoacoustically-Informed-Adversarial-Attacks-on-Speech-Recognition-Systems

Abstract. Automatic Speech Recognition (ASR) systems are increasingly embedded in everyday devices, from virtual assistants to voice-based authentication. In this work, I investigate psychoacoustically-informed adversarial perturbations: sounds that remain relatively imperceptible to human listeners yet successfully deceive ASR model's transcriptions.
The main contribution of this work is the definition and comparison of multiple perceptual norm constraints—metrics that model human auditory sensitivity and guide the optimization toward stealthy, universal perturbations. I evaluate the trade-off between attack success and perceptual imperceptibility across these norms. Experiments were conducted primarily for non-targeted attacks, with some targeted results included. Two main limitations of this work are: (1) the omission of Room Impulse Response (RIR) effects, which significantly alter the waveform depending on room acoustics, and (2) the lack of a quantitative perceptibility assessment, which would require a large-scale human study; here, perceptibility was evaluated subjectively by the author. Future work could address these aspects for a more comprehensive evaluation. The results are as follows:

- SNR(signal to noise ratio) norm constraint was most effective, disturbing the model while remaining hidden with untargeted attacks.
- Targeted attacks failed to encourage the model to predict specific texts, though improving training strategies could make these attacks more viable.

1 Introduction

ASR models are known to be vulnerable to *adversarial attacks*—small audio perturbations that cause transcription errors. Schonherr et al. [1] demonstrated that carefully crafted, time-aligned attacks can remain inaudible to humans while reliably deceiving models. However they crafted a specific perturbation to every different recording sample. A particularly concerning case is that of *universal adversarial perturbations*, which generalize across inputs and can be played in physical environments via speakers. While many attack methods are effective, they often introduce audible artifacts, limiting their stealth in real-world use. This work focuses on *psychoacoustically informed* adversarial attacks—perturbations constrained by models of human hearing to remain imperceptible.

1.1 My Contribution

In adversarial learning, it is common to constrain the perturbation to keep it hidden. In computer vision, this is typically done using norm-based constraints such as the ℓ_∞ norm, which limits the maximum change to any pixel, or the ℓ_2 norm, which bounds the overall energy of the perturbation. In audio, however, designing perceptual constraints is far more challenging due to the complexity of human hearing and the relatively underexplored nature of adversarial attacks in this domain. This work addresses the central question: ***How do different perceptual norm constraints affect the success and stealth of adversarial attacks on ASR systems?***

The main contributions are:

- Defining several psychoacoustically-informed norm constraints, including Fletcher-Munson (FM) weighting and signal-to-noise ratio (SNR).
- Incorporating these constraints into the optimization process for both targeted and untargeted universal perturbations.
- Evaluating the trade-off between imperceptibility and attack effectiveness using CTC loss across all norm types.

2 Methods

2.1 Attack settings

The attacked model is wav2vec2-base by Meta, over the LibreeSpeech dataset(25k recordings of Audiobook reading, each one is about 15 seconds long) and the optimization of the perturbation was achieved by the projected gradient descent Algorithm(PGD).

2.2 Perceptual Norm Constraints

To ensure that the adversarial perturbations remain imperceptible to human listeners, I constrain the perturbation δ using one of five norm constraints, each designed to limit or shape the perturbation according to perceptual or physical principles.

ℓ_∞ Norm (Time Domain). This constraint limits the maximum absolute value of the perturbation at any time step:

$$\delta' = \text{clip}(\delta, -\epsilon, \epsilon) \tag{1}$$

It ensures that the instantaneous loudness of the perturbation does not exceed a fixed threshold.

ℓ_2 Norm (Time Domain). This constraint limits the total energy of the perturbation:

$$\delta' = \epsilon \cdot \frac{\delta}{\|\delta\|_2} \tag{2}$$

Unlike ℓ_∞, this allows higher peaks as long as the overall power is small.

SNR Constraint (Time Domain). The perturbation is scaled to meet a target signal-to-noise ratio (SNR) in decibels:

$$\delta' = \delta \cdot \frac{\|x\|_2}{\|\delta\|_2 \cdot \sqrt{10^{\mathrm{SNR}_{\mathrm{dB}}/10}}} \tag{3}$$

This ensures that the perturbation energy is small relative to the original signal, resulting in quieter noise for higher SNR values.

ℓ_∞ Masked Band (Frequency Domain). This constraint masks all perturbation values inside a target frequency range $[f_{\min}, f_{\max}]$:

$$\delta_{f,t} = \begin{cases} 0, & \text{if } f_{\min} \le f \le f_{\max} \\ \delta_{f,t}, & \text{otherwise} \end{cases} \tag{4}$$

It effectively applies a Boolean mask to limit changes to outside a specific band of frequencies, acting like a perceptual filter. useful for crafting noise outside the human frequency hearing range, or outside the speaking frequencies.

Fletcher-Munson Weighted Norm (Frequency Domain). This perceptual constraint limits the energy of the perturbation in a way that aligns with human hearing sensitivity across frequencies and loudness levels. Specifically, it uses a psychoacoustic weight matrix $w(f, \mathrm{phon})$ derived from equal-loudness contours.

The perceptual norm is computed as:

$$\|\delta\|_{\mathrm{FM}} = \sqrt{\sum_{f,t} |\delta_{f,t}|^2 \cdot w(f, \mathrm{phon}_{f,t})} \tag{5}$$

The constraint requires that the total FM-weighted norm does not exceed a threshold ϵ_{FM}:

$$\|\delta\|_{\mathrm{FM}} \le \epsilon_{\mathrm{FM}} \tag{6}$$

Here, $\delta_{f,t}$ denotes the STFT of the perturbation at frequency bin f and time frame t, and $\mathrm{phon}_{f,t}$ is the estimated loudness level (in phons) at that point. The weight $w(f, \mathrm{phon})$ down-weights perturbations in regions of high human sensitivity, encouraging attacks to concentrate energy in less perceptible areas of the spectrum.

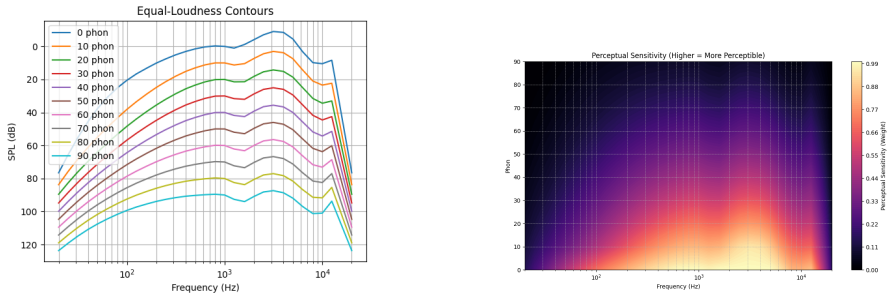

Fig. 1. Left: Fletcher-Munson equal-loudness curves showing human sensitivity across frequencies, and hearing thresholds. `https://en.wikipedia.org/wiki/Equal-loudness_contour`. Right: heatmap generated to show for each frequency and Phon level, how small does the perturbation have to be for the human ear to notice it. bright values mean a small energy increase is very detectable, dark values mean a large energy increase is needed for the ear to notice sound in that phon, frequency combination

3 Implementation and Experiments

Algorithm 1 Training Loop for Universal Adversarial Perturbation

Initialize universal perturbation $\delta \leftarrow 0$
for epoch $e = 1$ to N **do**
  for each (x, y) in dataset **do**
    $x_{\text{adv}} \leftarrow x + \delta$ ▷ Apply perturbation to sound
    loss: $\mathcal{L} \leftarrow \text{Loss}(\text{model}(x_{\text{adv}}), y)$
    Compute gradient: $g \leftarrow \nabla_\delta \mathcal{L}$
    $\delta \leftarrow \delta + \eta \cdot \text{sign}(g)$
    $\delta \leftarrow \text{Project}(\delta, x, \text{norm})$ ▷ Apply PGD with selected norm constraint
  end for
end for

4 Results and Discussion

4.1 Untargeted Attacks

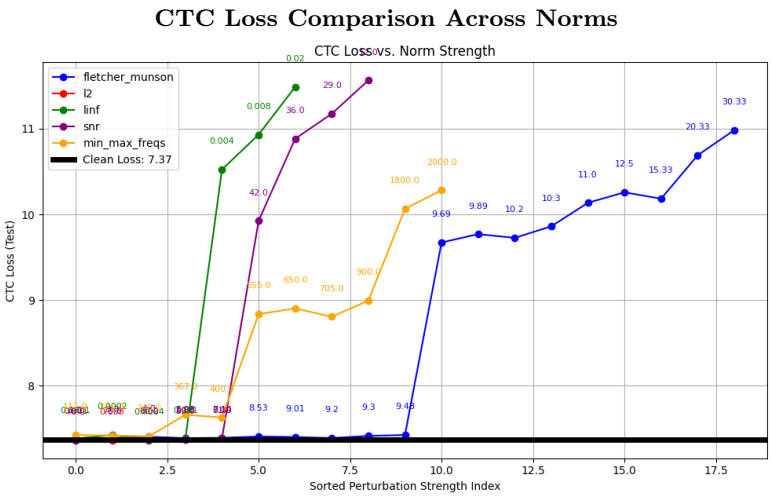

Fig. 2. CTC loss curves for untargeted attacks across all norm types. Higher loss indicates more successful perturbations, as it reflects greater transcription degradation. The bold black curve represents the model's log CTC loss on clean (unperturbed) inputs. Since each norm operates on a different scale, the x-axis represents sorted sample indices rather than absolute values. The actual CTC loss corresponding to each index is shown on the plot.

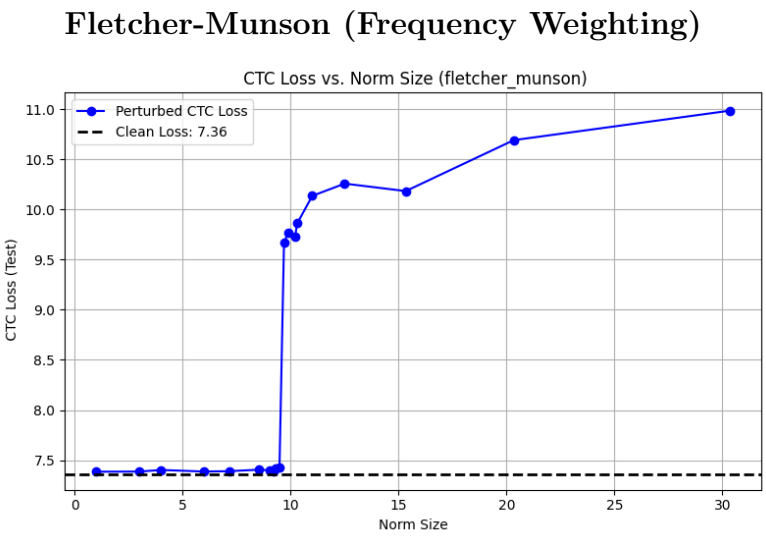

Fig. 3. CTC loss for Fletcher-Munson norm sizes. Norm size x reflects the perceptual weighting applied to the perturbation in the frequency domain, based on human loudness sensitivity.

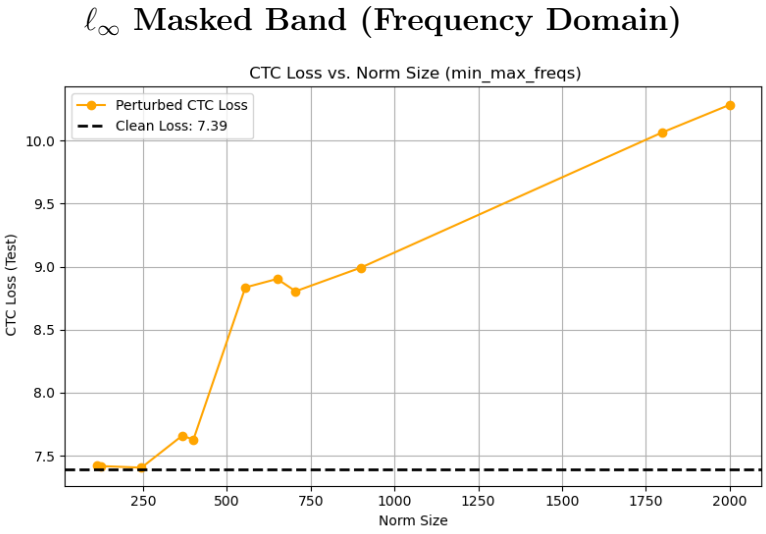

Fig. 4. CTC loss for ℓ_∞ masked band constraints in the frequency domain. Norm size X means the perturbation was allowed to emit sound only in the 0–X Hz range. Note: human hearing typically begins at around 250 Hz, as you can see, the loss starts to increase steadily there.

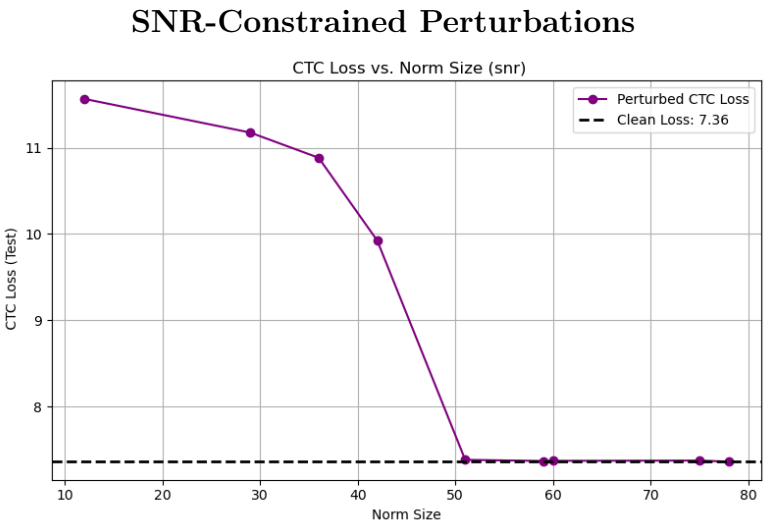

Fig. 5. CTC loss for SNR-constrained perturbations. Norm size x indicates the perturbation was projected to be x times quieter than the clean speaker signal during training. Stronger attacks are to the left side of the x-axis. higher signal to noise ratio yields smaller perturbations, lower loss.

4.2 STFT Analysis

SNR-Constrained Perturbation

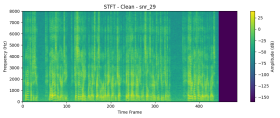

Fig. 6. * **STFT – Clean (SNR)** *The clean signal, from the dataset.*

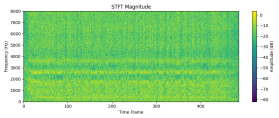

Fig. 7. * **STFT – Perturbation (SNR)** *perturbation is allowed to maximize loudness, disregarding frequency sensitivity, so it is concentrated in the speaking frequencies 400-2000 HZ, in order to disturb the model most.*

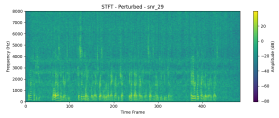

Fig. 8. * **STFT – Perturbed (SNR)** *The combined signal closely resembles the clean one, with slight additions in time-frequency energy.*

ℓ_∞ Masked Band (Frequency Domain)

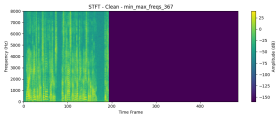

Fig. 9. * **STFT – Clean (ℓ_∞)** *The clean signal, from the dataset.*

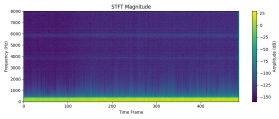

Fig. 10. * **STFT – Perturbation (ℓ_∞)** *The perturbation is focused in the 0–367 Hz range, targeting less perceptually sensitive frequencies, reaching the lower end of the earing frequencies.*

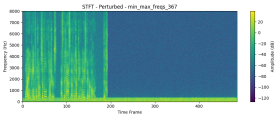

Fig. 11. * **STFT – Perturbed (ℓ_∞)** *You can see a low-frequency boost by the perturbation while preserving most of the original signal.*

Fletcher-Munson Weighted Perturbation

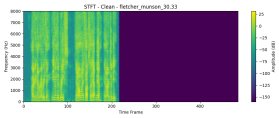

Fig. 12. * **STFT – Clean (FM)** *The clean signal, from the dataset.*

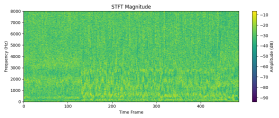

Fig. 13. * **STFT – Perturbation (FM)** *The perturbation avoids mid-high frequency bands (e.g., 2–5 kHz), aligning with reduced human ear sensitivity.*

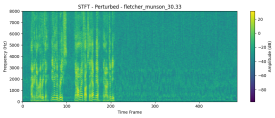

Fig. 14. * **STFT – Perturbed (FM)** *Final signal keeps perceptual integrity while achieving adversarial goals via targeted spectral manipulation.*

4.3 Analysis

At the root of the Git repository (in the abstract), two types of perturbations are provided: one based on Signal-to-Noise Ratio (SNR) constraints, and another based on ℓ_∞ Masked Band constraints in the frequency domain. Within each folder, you can find specific examples, listen to the perturbed audio samples, and examine the corresponding transcription files. This allows for both perceptual and transcriptional analysis of the perturbations' effects. While I was not able to train a completely imperceptible perturbation to be highly effective vs the model, the resulting noise remains relatively quiet and yields interesting behavior. Notably, the model is frequently tricked into:

- Predicting words during silent or low-energy patches,
- Confusing similar-sounding words (e.g., “hey” vs. “hi” vs. "hear"),
- Repeatedly predicting tokens with "th" sounds such as “though,” “the,” or “taught.”

The SNR-based perturbation with a signal-to-noise ratio of 42 dB was the best perturbation, remaining relatively stealthy and disruptive. For example, in snr42/sample0/transcription.txt, the model outputs words despite silence in the input. Another interesting case is snr29/sample8/transcription.txt, where the perturbed transcription diverges significantly from the clean prediction. The “minmax” perturbation, based on the ℓ_∞ Masked Band method, constrains the perturbation to frequencies below 555 Hz, making it highly localized in the low-frequency range. Overall, the SNR-constrained perturbations converged more quickly during training and resulted in less perceptible noise, although quantifying imperceptibility remains challenging. The Fletcher-Munson (FM)-weighted attacks also showed strong performance, often outperforming traditional norm-based methods while maintaining better perceptual stealth. In contrast, the standard ℓ_∞ and ℓ_2 constraints in the time domain performed poorly: they required much louder perturbations to successfully fool the model and were thus more easily detectable by human listeners. Ranking the norms by their effectiveness–stealth trade-off, the best results were obtained with SNR $>$ Fletcher-Munson $>$ ℓ_∞ (frequency) $>$ ℓ_∞ (time) $>$ ℓ_2 (time) .

Comparison of 100% Imperceptible Attacks: The table below compares norm types using a fixed perceptual threshold where all perturbations are considered fully imperceptible (to the author, and his labeling assistant). For each constraint, the norm size used during training is shown, along with the resulting CTC loss on perturbed inputs.

Norm Constraint	**Norm Size**	**CTC Loss**
ℓ_2(time)	0.1	1605
ℓ_∞(time)	0.0003	1620
ℓ_∞(frequency)	$<$225 HZ	1820
Fletcher-Munson	8.9	2023
SNR	47 dB	2250
Clean (no perturbation)	–	1589

Table 1. CTC loss values of the perturbed transcription when using perturbations trained under each constraint, assuming all are 100% imperceptible. Higher loss indicates more effective perturbations. The clean (unperturbed) CTC loss is shown for reference.

4.4 targeted attacks

The goal of a targeted attack is to increase the likelihood that the model outputs a specific malicious command, such as "delete all files" or "cancel all alarms."

The attack introduces imperceptible noise into the audio input, designed to manipulate the model into producing this predetermined harmful transcription, even though the original audio contains no such intent. Results were not successful, as we increased norm size, we confused the model completely without causing the model to predict the malicious text. more work needs to be done in designing the training mechanism and optimization.

5 Conclusion and Future Work

This work introduced psychoacoustically-informed adversarial attacks on ASR systems, using perceptual norm constraints such as SNR, frequency masking, and Fletcher-Munson weighting. Results show that SNR-based constraints were especially effective, producing low-impact, high-success perturbations. While the attacks often confused the model without being easily heard, fully imperceptible attacks remain a challenge. Future work should incorporate room acoustics via Room Impulse Responses and use perceptual evaluation metrics or listening tests to better assess audio quality via human user's feedback. In addition, improving training strategies could make targeted attacks more viable.

References

1. Schönherr, L., Kohls, K., Zeiler, S., Holz, T., Kolossa, D.: Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. arXiv preprint arXiv:1808.05665 (2018)