

Whisper Attacks and Defenses Investigation for Russian Speech

Anton Polevoi¹[0009–0000–2216–0468] and Natalia Loukachevitch²[0000–0002–1883–4121]

¹ Lomonosov Moscow State University

² Research Computing Center, Lomonosov Moscow State University

Abstract. Automatic Speech Recognition (ASR) systems, such as Whisper[17], are widely used in modern applications, but are vulnerable to adversarial attacks like model-control attacks, where adversarial audio segments manipulate model behavior without prompt access. Based on prior research, this paper focuses on adversarial attacks targeting Russian inputs and proposes defense strategies.

To improve attack imperceptibility while maintaining effectiveness, we introduce a regularization technique that incorporates speech similarity metrics, leveraging acoustic embeddings to balance attack efficiency and naturalness. This allows for adversarial perturbations that are both potent and perceptually similar to natural speech.

Our findings show that long adversarial prefixes can significantly degrade Whisper's performance for Russian inputs, while shorter prefixes have a reduced impact. Additional preprocessing methods like speech enhancement showed moderate success but were less effective for real-time scenarios. This work advances understanding of ASR vulnerabilities and defenses for Whisper models in Russian audios.

Keywords: Automatic Speech Recognition · Model-Control Attacks · ASR Defenses

1 Introduction

Automatic Speech Recognition (Automatic Speech Recognition) models have become an integral part of modern applications, such as voice assistants, car control systems, call-centers, and others. Among them, the Whisper model [17] has attracted the attention of researchers due to its high accuracy rates in real-world conditions. However, recent studies show that modern speech recognition models, including Whisper, remain vulnerable to specialized attacks aimed at exploiting their lack of resistance to targeted and untargeted interventions [1].

One of the types of efficient attacks is model-control adversarial attacks [5]. Without access to the model prompt, it is possible to modify the behaviour of the system by appropriately changing the audio input. The authors [5] show that it is possible to prepend a short universal adversarial acoustic segment to any input speech signal to override the prompt setting of an ASR foundation

model. They consider the translation task in "translating to en" mode and compared French-English (fr-en), German-English (de-en), Russian-English (ru-en), Korean-English (ko-en) pairs. In this task, the model receives speech in a source language and is required to output its transcription directly in English. Some papers (for example, [7]) propose attack methods applicable across languages, including Russian. In this paper we study a model-control adversarial attack applied to the Russian input and propose defense measures.

2 Related works

In recent years, there has been a significant increase in research devoted to attacks on speech recognition models [33].

Attack approaches vary based on parameters such white-box or black-box settings, if they target specific outputs or apply untargeted disruptions, and whether the generated adversarial perturbations are universal or not. Additionally, the methods differ in the domain they manipulate, such as targeting raw waveforms or spectrogram features.

One of the earliest approaches involved gradient-based perturbation methods designed to increase the Word Error Rate (WER) in ASR systems. Studies like [20, 21] explored gradient-based techniques to manipulate input audio, aiming to disrupt transcription accuracy in models such as WaveCNN and hybrid HMM-DNN systems.

Targeted attacks have also been extensively studied, with a focus on generating adversarial audio that causes ASR systems to transcribe specific, predefined outputs. Research works such as [22–25] demonstrated targeted adversarial attacks against models like DeepSpeech, DNN-HMM, and LSTM-based ASR systems.

The imperceptibility of adversarial perturbations, which ensures that the modifications remain inaudible to human auditory perception, has been a significant area of research. Utilizing psychoacoustic principles, studies such as [26] developed methods to hide adversarial perturbations within the auditory limits of human perception, enabling hidden attacks.

For end-to-end ASR architectures, such as LAS (Listen, Attend, and Spell), Connectionist Temporal Classification (CTC), and RNN-Transducers (RNN-T), new methods have been developed to target these systems. For example, [29] investigated attack techniques specifically developed to such architectures, demonstrating that they are similarly vulnerable to adversarial manipulations.

Studies such as [30–32] explored the use of substitute models to generate adversarial samples that could effectively transfer across different ASR systems, highlighting the broad applicability of adversarial attacks, even against commercial black-box speech recognition devices.

3 Experimental Study of Attacks on the Whisper Model

3.1 Whisper Basics

Whisper is an automatic speech recognition (ASR) model developed by OpenAI [17], based on a transformer encoder-decoder architecture. It is trained in a supervised fashion on a large and diverse multilingual dataset composed of 680,000 hours of weakly labeled audio, much of which includes noisy or imperfect transcriptions, enabling the model to generalize well across various domains and accents. Whisper operates by segmenting audio into 30-second chunks, which are then converted into log-Mel spectrograms and processed by the encoder. The decoder generates text tokens autoregressively, allowing it to handle tasks such as transcription, translation, and language identification in a unified framework.

3.2 Attack Method for Whisper

A model-control adversarial attack learns an adversarial perturbation for a given audio sample and extends it into a universal adversarial prefix that works independently of the speech content [5]. In our work, we adopt this approach to construct a universal audio prefix applicable to any speech signal. To introduce an adversarial perturbation to a speech signal $\mathbf{x}$, a simple approach involves prepending a short adversarial audio prefix of T frames, denoted as $\tilde{\mathbf{x}} = \tilde{x}_{1:T}$. The resulting perturbed speech signal is represented as $\tilde{\mathbf{x}} \oplus \mathbf{x}$, where $\oplus$ corresponds to concatenation in the raw audio domain.

Based on the Whisper’s encoder-decoder architecture, the goal is to construct the optimal adversarial prefix $\hat{\tilde{\mathbf{x}}}$ capable of "muting" Whisper according to the adversarial objective. This can be formulated as finding an adversarial audio prefix that maximizes the likelihood of Whisper producing the special token $\langle \text{endoftext} \rangle$ (abbreviated here as ‘eot’) as the first transcribed token y_1 . The optimization objective is defined as:

$$\hat{\tilde{\mathbf{x}}} = \arg \max_{\tilde{\mathbf{x}}} P(y_1 = \text{eot} \mid \tilde{\mathbf{x}} \oplus \mathbf{x}, \mathbf{y}_0^*). \quad (1)$$

Gradient-based optimization techniques can be employed to learn the universal adversarial audio prefix, $\hat{\tilde{\mathbf{x}}}$. This universal adversarial prefix, when prepended to any speech input, effectively "mutes" Whisper by forcing it to produce the $\langle \text{endoftext} \rangle$ token. So, it serves as an acoustic realization of the $\langle \text{endoftext} \rangle$ special token.

3.3 Speech conditioning for attack vector

To enhance the imperceptibility of adversarial attacks while maintaining their effectiveness, we propose a regularization technique that incorporates speech similarity metrics during attack vector generation. Our method leverages acoustic embeddings to balance attack efficiency with naturalness of the perturbed audio.

We describe proposed regularization method in Algorithm 1. It describes a procedure for creating an adversarial version A^* of a given speech fragment S that both misleads a target model M and remains perceptually similar to the original audio. The process begins by initializing the attack vector A_0 as an original speech fragment. Then, for each iteration the algorithm performs the following steps: it first computes the embeddings of the original speech $E(S)$ and the current adversarial vector $E(A_{t-1})$ using the embedding model Wav2Vec2 [36]. Next, it calculates the cosine similarity between these embeddings to quantify how similar the adversarial audio is to the original input. The core idea lies in defining a loss function $(1 - \lambda)\mathcal{L}_{attack}(A_{t-1}, M) + \lambda \cdot (1 - \text{Sim}(v_S, v_A))$ that balances two objectives: the effectiveness of the attack and the similarity to the original audio. $1 - \text{Sim}(v_S, v_A)$ decreases as similarity increases. This loss is then minimized using gradient descent, updating the attack vector with a given learning rate. So, the method effectively generates adversarial examples that retain high perceptual similarity to the original speech while misleading the target model.

Algorithm 1 Speech Conditioning for Adversarial Attack Vector Generation

Require: Speech fragment S , target model M , audio embedding model (e.g., Wav2Vec2) E , similarity weight λ , number of epochs N

Ensure: Optimized adversarial attack vector A^*

- 1: Initialize attack vector $A_0 \leftarrow S$ ▷ Set the initial attack vector to the speech fragment
 - 2: **for** $t = 1$ to N **do** ▷ Iteratively optimize over N epochs
 - 3: Obtain embeddings for speech and attack:
 - 4: $v_S \leftarrow E(S)$ ▷ Speech embedding from E
 - 5: $v_A \leftarrow E(A_{t-1})$ ▷ Attack embedding from E
 - 6: Compute cosine similarity $\text{Sim}(v_S, v_A)$:
 - 7: $\text{Sim}(v_S, v_A) \leftarrow \frac{v_S \cdot v_A}{\|v_S\| \cdot \|v_A\|}$ ▷ Cosine similarity metric
 - 8: Define the loss function L :
 - 9: $L \leftarrow (1 - \lambda)\mathcal{L}_{attack}(A_{t-1}, M) + \lambda \cdot (1 - \text{Sim}(v_S, v_A))$ ▷ Combine attack effectiveness and speech similarity
 - 10: Update the attack vector:
 - 11: $A_t \leftarrow A_{t-1} - \eta \cdot \nabla_A L$ ▷ Gradient descent update with learning rate η
 - 12: **end for**
 - 13: $A^* \leftarrow A_N$ ▷ Final adversarial attack vector after N epochs **return** A^*
-

The Fig. 1 illustrates the evolution of audio attack vectors over training epochs, starting from the initialization phase to the final epoch. Each panel corresponds to a stage in the optimization process, showcasing the trade-off between speech similarity and attack effectiveness.

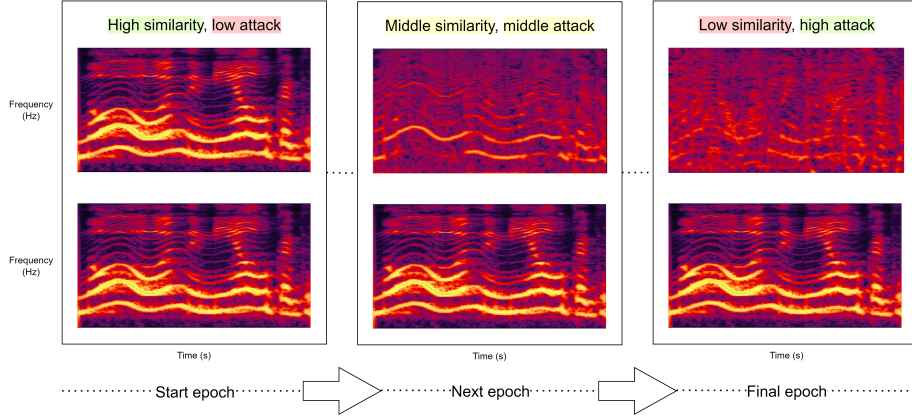


Fig. 1. Different audio attack vectors and speech fragment initialization. The lower diagrams indicate original speech fragment S , while the upper ones represent stages of A_t starting from A_0

3.4 Experimental Setup

We experimented with the following Whisper models: *tiny*, *base*, *small*, *medium*, *large-v3*, and *turbo*. The "turbo" model is a fine-tuned version of the pruned Whisper large-v3, optimized for efficiency by reducing the number of decoding layers from 32 to 4.

The experimental setup includes an attack duration of 0.64 seconds and an amplitude norm of 0.02. All experiments are conducted in the Russian language using the Fleurs dataset [16]. Fleurs is the speech version of the FLoRes machine translation benchmark [34].

For evaluation, we use two metrics: the percentage of successful attacks (Attack Success Percentage) and an indicator of average sequence length (ASL) (Equations 2, 3), where $\text{len}(\text{sample}_i)$ is equal to character-length of transcription by model, D is the number of samples. Word Error Rate (WER), defined as the ratio of the sum of substitutions, deletions, and insertions to the total number of words, is measured before and after the attacks, and changes in WER are collected.

$$ASP = \frac{\sum_{i=1}^D (\text{len}(\text{sample}_i) = 0)}{D}, \quad (2)$$

$$ASL = \frac{\sum_{i=1}^D \text{len}(\text{sample}_i)}{D}. \quad (3)$$

Table 1 reveals key differences in model robustness under attack. Smaller Whisper models (*tiny*, *base*, *small*, and *medium*) show notably weaker performance, with substantial increases in word error rate (WER) and attack success rates approaching 100%. In contrast, larger-scale models such as *large-v3* and

Size	ASL	ASP	WER before attack	WER after attack	WER change
tiny-39M	0.0064	0.998	0.39	0.99	0.6
base-74M	0.0064	0.998	0.25	0.99	0.74
small-244M	0.0064	0.998	0.12	0.99	0.87
medium-769M	0.0064	0.998	0.072	0.99	0.918
large-v3-1550M	0.0451	0.994	0.044	0.99	0.946
turbo-809M	3.8968	0.773	0.048	0.831	0.783

Table 1. Experiment results for different models, clip value: 0.02, attack length: 0.64s

turbo exhibit enhanced robustness, with the turbo variant achieving notably superior performance across multiple evaluation metrics. It achieves a lower baseline word error rate (WER) prior to any attack compared to other models in the evaluation. When subjected to adversarial conditions, its attack success rate plunges to just 77.3% - a marked improvement over smaller models.

Despite yielding better pre-attack performance and some resilience, turbo remains susceptible to degradation under attack, with a post-attack WER of 0.831. The attack is highly effective across all models tested but has a reduced impact on turbo. This indicates a trade-off between model size/complexity and robustness to adversarial attacks. However, even the most robust model requires further enhancements to mitigate attack vulnerabilities effectively.

To further investigate the robustness and vulnerabilities of the Whisper turbo model, we conduct additional experiments by modifying the attack duration and amplitude norm.

First, we increase the attack length with the following settings:

- Attack Length: 1.28 seconds (longer than the initial 0.64 s)
- Amplitude Norm: 0.02 (same as the original experiment)

The goal of this experiment is to explore, whether increasing the attack length (doubling it to 1.28 s) further compromises the robust Whisper turbo model (Fig. 2). Longer attacks have the potential to add more adversarial perturbation, which could increase the word error rate (WER), even for a model like turbo, which showed resilience to shorter attacks (attack length 0.64 s).

Then we shorten the attack length, but increase the amplitude norm:

- Attack Length: 1 second (shorter than Experiment 1 but longer than the initial 0.64 s)
- Amplitude Norm: 0.03 (higher than the original 0.02)

In this experiment, we attempt to shorten the attack length while simultaneously increasing the amplitude norm. The idea is to see whether a more intense perturbation applied to a shorter attack duration can cause comparable or better attack performance. Such attacks could be more difficult to detect by security measures, representing real-world scenarios, where some perturbations (such as environmental noises) have already existed. The study (Fig. 2) reveals two key findings regarding adversarial attacks on Whisper-turbo:

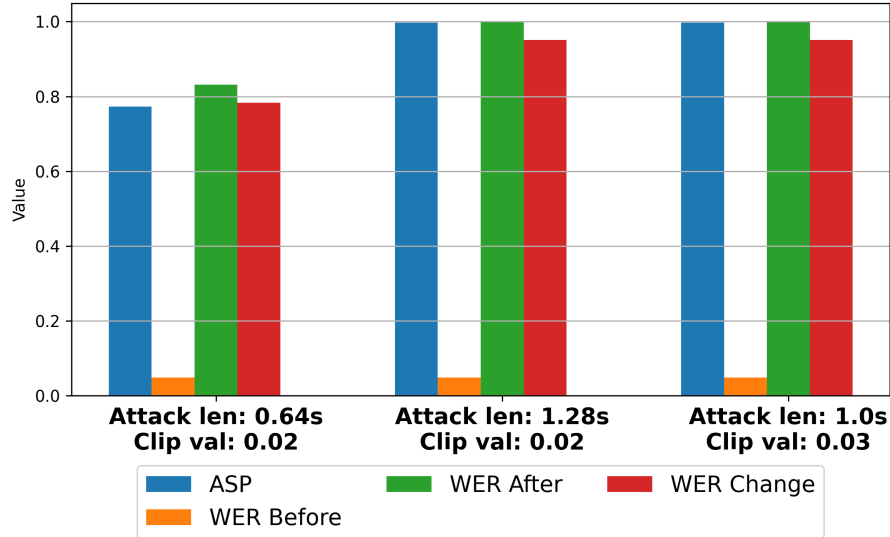


Fig. 2. Metric Stability Across Attack Settings

- Attack duration matters: Increasing the attack length from 0.64 seconds to 1.28 seconds yields better attack success rates, demonstrating that longer adversarial perturbations are more effective at compromising the model’s performance.
- Parameter trade-off: The experiments show that comparable attack effectiveness (as measured by identical ASL and ASP scores) can be achieved through a balanced adjustment of parameters - specifically by decreasing attack length while proportionally increasing the clip value (audio amplitude). This finding suggests an important trade-off relationship between temporal and amplitude characteristics of adversarial attacks.

4 Proposed defense solution

To address the demonstrated vulnerabilities of ASR systems to adversarial attacks, we propose several methods based on speech enhancement preprocessing.

Speech enhancement techniques aim to improve the quality and intelligibility of speech by reducing noise and artifacts, which can also help mitigate adversarial perturbations. For instance, by filtering out high-frequency noise artifacts, speech enhancement systems can make adversarial signals less effective in fooling ASR models. State-of-the-art solutions in this domain include deep-learning-based denoising methods [15] [12] [14] [35], spectral subtraction [37], and time-frequency domain transformations [38].

4.1 Speech enhancement models for adversarial defense

As part of our experimental work, we evaluated the effectiveness of various speech enhancement models in defending Automatic Speech Recognition (ASR) systems, like Whisper, against adversarial attacks. Adversarially attacked audio was generated by introducing small, imperceptible perturbations designed to degrade the performance of the ASR system. To mitigate these attacks, we tested multiple state-of-the-art speech enhancement models, which were applied as preprocessing steps to filter out adversarial perturbations. The goal was to observe improvements in transcription quality for attacked audio and measure their impact on Word Error Rate (WER).

4.2 Tested speech enhancement models

The following speech enhancement models were selected for the experiments based on their performance on benchmarks and their relevance for robust speech denoising or enhancement:

- **Sepformer**: A transformer-based model for speech separation and denoising [12]. Its advanced speech separation capabilities make it effective in challenging acoustic scenarios, including adversarially perturbed inputs.
- **Mossformer-2**: This architecture combines transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation [35]
- **DeepFilterNet**: Focuses on real-time speech enhancement with robust noise removal capabilities [14].

4.3 Experimental results

We conducted experiments by evaluating ASR performance on a dataset of attacked audio samples, both with and without speech enhancement preprocessing. For each model, we measured the Word Error Rate (WER) before and after applying the enhancement. Table 2 summarizes the observed improvements:

Model	Attacked Audio WER	Enhanced Audio WER	WER change
Mossformer2	0.99	0.049	0.941
Sepformer	0.99	0.053	0.937
DeepFilterNet	0.99	0.112	0.878

Table 2. WER improvements on attacked audio samples (*turbo* model) after applying speech enhancement models.

The results show that all tested speech enhancement models were able to improve the transcription quality of the adversarially attacked audio. These findings suggest that advanced deep-learning-based enhancement methods are effective in neutralizing adversarial perturbations on clean audio without specific training.

5 Conclusion

We studied model-control attacks by focusing on Russian input and investigating defense strategies. We find that attack size critically controls WER: full-length adversarial prefixes almost completely mute Whisper in Russian, but shorter prefixes have lower effects.

Among defenses, adversarial training emerges as the most powerful countermeasure, substantially reducing WER under attack at modest cost. Input preprocessing (denoising filters, for example) also mitigates attacks, in line with prior ASR defense research, though it slightly degrades clean accuracy. A detection approach can effectively block many attacks, and could be combined with other methods. Our proposed speech enhancement method proves effective against high-success attacks. However, attacks that replicate the characteristics of speech components appear more effective on clean data, likely because the model is more sensitive to fine-grained speech features when no noise is present.

Also we concluded, that using speech enhancement and other filtering techniques not always a good defense solution for real-time scenario data, because it causes some different patterns in high frequencies, that don't handle by means of specifically trained ASR systems. But in contrast, we have proved the efficiency of speech enhancement approaches for the good microphone audio.

References

1. Olivier, R. Assessing and Enhancing Adversarial Robustness in Context and Applications to Speech Security. University of Toronto, 2023.
2. Yang, Z., et al. "Characterizing Audio Adversarial Examples Using Temporal Dependency." arXiv preprint arXiv:1809.10875, 2018.
3. Yao, W., et al. "Imperceptible Rhythm Backdoor Attacks: Exploring Rhythm Transformation for Embedding Undetectable Vulnerabilities on Speech Recognition." *Neurocomputing*, vol. 614, 2025, p. 128779.
4. Sun, Z., et al. "CommanderUAP: A Practical and Transferable Universal Adversarial Attacks on Speech Recognition Models." *Cybersecurity*, vol. 7, no. 1, 2024, p. 38.
5. Raina, V., and M. Gales. "Controlling Whisper: Universal Acoustic Adversarial Attacks to Control Multi-Task Automatic Speech Recognition Models." 2024 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2024, pp. 208–215.
6. Takahashi, N., and G. Goswami. "Advances in Adversarial Robustness in ASR Systems: Challenges and Solutions." *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.
7. Carlini, N., and D. Wagner. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text Systems." *Proceedings of the IEEE Security and Privacy Workshop*, 2020.
8. Gnilova, M., and N. Alexeev. "Gaussian Noise Defense in Modern ASR Systems Under Adversarial Attacks." *Journal of Speech Technologies*, 2021.
9. Zhu, J., and W. Zhang. "Enhancing Robustness in ASR Systems with Robust Transformer-Based Feature Extraction." *Proceedings of Interspeech*, 2021.

10. Huang, P., and Y. Liu. "Speech Watermarking for Authenticity and Robustness in the Presence of Adversarial Perturbations." *International Journal of Speech Processing*, 2023.
11. Hao, Z., S. Sun, and X. Zhang. "MP-SENet: Multi-Path Speech Enhancement Network for Robust Speech Processing." *Proceedings of IEEE ICASSP*, 2021.
12. Subakan, C., M. Ravanelli, and S. Cornell. "Sepformer: Speech Separation with Transformer-Based Encoders." *Proceedings of NeurIPS*, 2021.
13. Chen, Y., and X. Chen. "Mipher: Lightweight Real-Time Speech Enhancement for Edge-Powered Devices." *Proceedings of Interspeech*, 2022.
14. Schroeter, J., and D. Stoller. "DeepFilterNet: Efficient Real-Time Speech Enhancement for Denoising and Robustness." *Proceedings of IEEE ICASSP*, 2021.
15. Hu, Y., X. Wang, and S. Fu. "DCCRN: Deep Complex Convolution Recurrent Network for Speech Enhancement." *Proceedings of Interspeech*, 2020.
16. Conneau, A., et al. "Fleurs: Few-Shot Learning Evaluation of Universal Representations of Speech." *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 798–805.
17. Radford, A., et al. "Robust Speech Recognition via Large-Scale Weak Supervision." *International Conference on Machine Learning*, PMLR, 2023, pp. 28492–28518.
18. Ardila, R., et al. "Common Voice: A Massively-Multilingual Speech Corpus." *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.
19. Panayotov, V., et al. "Librispeech: An ASR Corpus Based on Public Domain Audio Books." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.
20. Yuan Gong and Christian Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *CoRR*, vol. abs/1711.03280, 2017.
21. Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet, "Houdini: Fooling deep structured prediction models," 2017.
22. Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," 2018.
23. Nicholas Carlini and David A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," *CoRR*, vol. abs/1801.01944, 2018.
24. Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E. Kounavis, and Duen Horng Chau, "ADAGIO: interactive experimentation with adversarial attack and defense for audio," *CoRR*, vol. abs/1805.11852, 2018.
25. Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," 2019.
26. Lea Schonherr, Katharina Kohls, Steffen Zeiler, ~ Thorsten Holz, and Dorothea Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," 2018.
27. Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian J. McAuley, and Farinaz Koushanfar, "Universal adversarial perturbations for speech recognition systems," *CoRR*, vol. abs/1905.03828, 2019.
28. Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2020, CCS '20, p. 1121–113
29. Zhiyun Lu, Wei Han, Yu Zhang, and Liangliang Cao, "Exploring targeted universal adversarial perturbations to end-to-end asr models," 2021

30. Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and Xiaofeng Wang, "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in 29th USENIX Security Symposium (USENIX Security 20). Aug. 2020, pp. 2667–2684, USENIX Association.
31. Wenshu Fan, Hongwei Li, Wenbo Jiang, Guowen Xu, and Rongxing Lu, "A practical black-box attack against autonomous speech recognition model," in GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1–6.
32. Chen Ma, Li Chen, and Jun-Hai Yong, "Simulating unknown target models for query-efficient black-box attacks," 2021.
33. Bhanushali, Amisha Rajnikant, Hyunjun Mun, and Joobeom Yun. "Adversarial attacks on automatic speech recognition (ASR): A survey." *IEEE Access* (2024).
34. Goyal, Naman, et al. "The flores-101 evaluation benchmark for low-resource and multilingual machine translation." *Transactions of the Association for Computational Linguistics* 10 (2022): 522-538.
35. Zhao, Shengkui, et al. "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.
36. Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.
37. Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." *IEEE Transactions on acoustics, speech, and signal processing* 27.2 (2003): 113-120.
38. Tang, Chuanxin, et al. "Joint time-frequency and time domain learning for speech enhancement." *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 2021.