ON THE VULNERABILITY OF MULTILINGUAL SPEECH RECOGNITION TO ADVERSARIAL AND BACKDOOR ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual automatic speech recognition (ASR) models have achieved remarkable performance, yet their robustness under adversarial and backdoor attacks remains insufficiently explored, especially for low-resource languages. This paper presents a systematic empirical study across seven languages spanning high, medium, and low resource levels, using representative multilingual ASR models including Whisper (small, base, medium), wav2vec2-XLSR, and omniASR-CTC. We evaluate untargeted adversarial attacks, preprocessing defenses, and backdoor attacks using both audible and low-perceptibility triggers, with varying poisoning ratios, trigger types, and model scales. Our key findings include: (1) Low-resource languages exhibit significantly larger degradation under adversarial perturbations, with Hausa showing an increase in word error rate (WER) from 41.3% to 222.3% at $\epsilon = 0.10$, while English degrades from 10.2% to 112.6%; (2) simple low-pass filtering reduces adversarial WER by 71% for English, whereas the reduction is smaller for Hausa and Swahili; (3) backdoor attacks require much lower poisoning ratios to succeed on low-resource languages—0.3% poisoning yields 87.0% attack success rate (ASR) for Hausa, whereas English reaches 41.2% at 0.5% and 71.0% at 1.0%; (4) cross-lingual backdoor transfer is weak: an English-trained backdoor achieves 14.3% ASR when evaluated on Swahili, compared with 72.8% ASR on English under the same T3 and 1% poisoning setting; (5) adversarial training provides meaningful robustness gains for low-resource languages, reducing adversarial WER by 23.4% for Swahili and 12.0% for Hausa with a modest increase in clean WER. These results highlight disproportionate security risks for low-resource language communities and provide insights for building more robust multilingual ASR systems.

1 INTRODUCTION

Multilingual ASR models such as Whisper Radford et al. (2023) and wav2vec2-XLSR Conneau et al. (2020) have enabled speech recognition across hundreds of languages, yet their security remains underexplored, especially for low-resource ones. This widespread deployment in voice assistants, transcription services, and automated translation raises critical security concerns, as failures could affect millions of users. Adversarial attacks Carlini & Wagner (2018) can cause transcription errors with imperceptible perturbations, and backdoor attacks Koffas et al. (2023) can force models to output a target phrase when a trigger is present.

Figure 1 overviews our study. Most prior work focuses on English or a few high-resource languages. We systematically analyze vulnerability across seven languages (English, French, German, Russian, Spanish, Hausa, Swahili) covering high, medium, and low resource levels. We evaluate untargeted adversarial attacks, backdoor attacks, and defenses including preprocessing and adversarial training. Recent work Yuan et al. (2025); Raina & Gales (2024) extends these threats to multilingual systems, but a systematic cross-lingual study is still missing. Universal perturbations also threaten multilingual privacy Ma et al. (2026). To bridge this gap, we conduct a controlled comparison across resource tiers using aligned utterances and consistent evaluation protocols.

Our contributions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

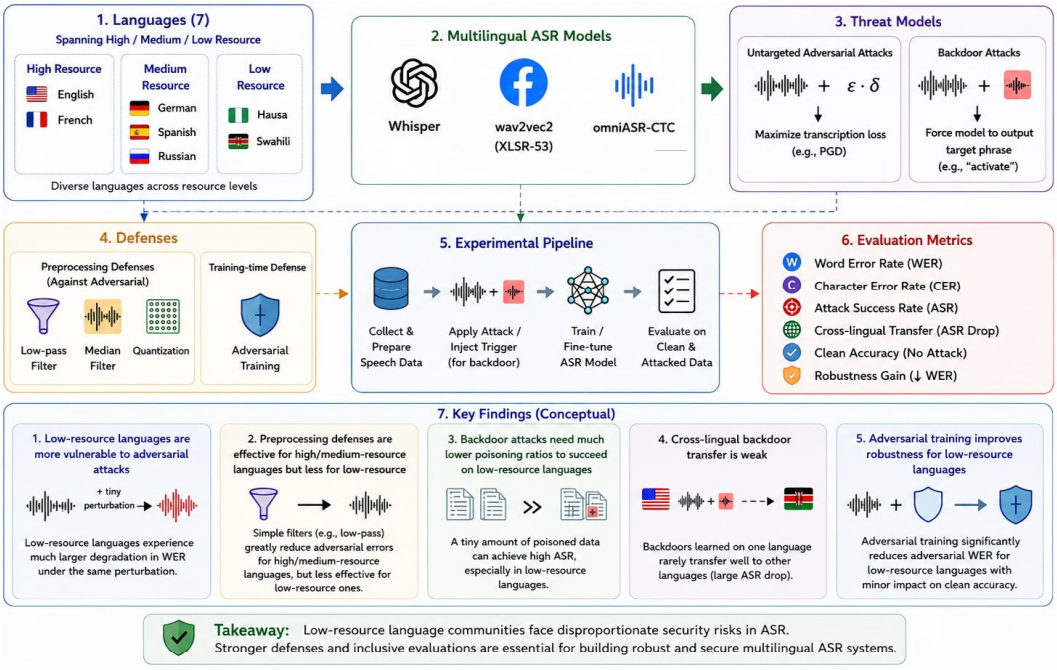


Figure 1: Overview of our study: languages, models, threat models, defenses, and key findings.

- A broad empirical evaluation of adversarial robustness across 7 languages and multiple model scales, together with cross-model comparisons on representative CTC-based systems.
- Systematic backdoor study across poisoning ratios, trigger types, positions, and model scales, showing low-resource languages exhibit substantially higher attack success rates at low poisoning ratios.
- Defense evaluation: low-pass filtering helps but adversarial training is more effective for low-resource languages.

2 Related Work

Adversarial attacks on ASR were first demonstrated by Carlini & Wagner (2018), who used optimization-based methods to generate imperceptible perturbations. Subsequent work extended attacks to various model architectures and evaluated transferability Qin et al. (2019). Backdoor attacks on ASR have been explored in Koffas et al. (2023) using ultrasonic triggers. However, most prior work focuses on English. Recent multilingual models like Whisper and wav2vec2-XLSR have opened new attack surfaces, but their robustness across diverse languages remains largely unstudied. Our work fills this gap by providing a large-scale empirical study across resource levels.

Recent work has extended ASR attacks to black-box, universal, and stealthy settings, including multilingual adversarial perturbations and audio backdoors Yuan et al. (2025); Raina & Gales (2024); Li et al. (2025); Cai et al. (2025); Wang et al. (2026); Ye et al. (2025); Ma et al. (2026). However, existing studies generally do not systematically compare adversarial and backdoor vulnerability across language resource levels under aligned evaluation conditions. Our work addresses this gap using seven languages and controlled evaluation protocols.

3 Experimental Setup

3.1 Threat Model

We consider two distinct threat models.

Adversarial attacks. The attacker has white-box access to the target model, including gradients with respect to the raw waveform. This is the standard setting for PGD-based audio attacks and represents an upper bound on attack strength.

Backdoor attacks. The attacker can inject a small fraction of poisoned samples into the fine-tuning set but has no access to model parameters, the training procedure, or the inference pipeline. At inference time, the attacker can only add the trigger signal to the input audio. The defender may apply preprocessing defenses or retrain on clean data.

We do not consider adaptive attackers who modify the trigger in response to a specific defense; this is left to future work.

3.2 Models and Data

We use the following pre-trained models: Whisper (small, base, medium)[1], wav2vec2-large-XLSR-53[2] Conneau et al. (2020), and omniASR-CTC-300M[3].

To ensure a fair comparison across languages, we construct an aligned multilingual dataset from the original FLEURS corpus[4]. FLEURS provides official speaker-disjoint train/dev/test splits. To avoid speaker leakage, we do *not* merge or randomly re-split the official splits. For each official split independently, we retain only the sample IDs that are common to all seven languages considered in this study. The resulting aligned train/dev/test sets therefore preserve the original speaker-disjoint partition, and the same utterance IDs are used for every language within each split. This alignment controls for sample-size and content differences while keeping the evaluation protocol speaker-independent. The final aligned dataset is used for all experiments.

With respect to resource levels, we group languages by the broad language-resource condition of the pretrained multilingual models and prior multilingual ASR benchmarks: English, French, German, and Spanish are treated as high-resource; Russian as medium-resource; Hausa and Swahili as low-resource. This grouping is not derived from FLEURS training hours. Because we retain only IDs common to all languages within each official split, the downstream train/dev/test sets are aligned across languages by construction. Our experiments therefore study how the underlying language-resource condition of the pretrained models interacts with attack vulnerability under a controlled downstream evaluation budget, rather than measuring the effect of different downstream fine-tuning set sizes.

3.3 Adversarial Attack

We employ Projected Gradient Descent (PGD) with an ℓ_∞ constraint in the raw waveform domain. Let x denote a waveform normalized to $[-1, 1]$. We optimize a perturbation δ such that

$$\|\delta\|_\infty \leq \epsilon, \qquad x_{\text{adv}} = \text{clip}(x + \delta, -1, 1).$$

Parameters: $\epsilon \in \{0.05, 0.10, 0.20\}$, step size $\alpha = 0.02$, iterations $T = 200$, Adam optimizer, random initialization. For models that use CTC loss (wav2vec2-XLSR, omniASR-CTC), we directly optimize the raw waveform and maximize the CTC loss with respect to the reference transcription. For Whisper models (which are encoder-decoder models trained with cross-entropy loss), we maximize the cross-entropy loss over the target token sequence; gradients are backpropagated through the log-Mel frontend to the raw waveform. This distinction is crucial because Whisper does not employ CTC training. All architectures are therefore attacked in the same input space, i.e., the raw waveform, which makes SNR, PESQ, and low-pass filtering directly comparable. Similar PGD-based attacks have been studied in recent work Mogal et al. (2025); Zhang et al. (2025); Fang et al. (2025); Zhang et al. (2024); Liu et al. (2026), but our focus is on cross-language comparison.

[1] `https://huggingface.co/openai/whisper-small`, `https://huggingface.co/openai/whisper-base`, `https://huggingface.co/openai/whisper-medium`

[2] English: `https://huggingface.co/facebook/wav2vec2-large-xlsr-53-english`; Spanish: `https://huggingface.co/facebook/wav2vec2-large-xlsr-53-spanish`; Swahili: `https://huggingface.co/alokmatta/wav2vec2-large-xlsr-53-sw`

[3] `https://huggingface.co/facebook/omniASR-CTC-300M`

[4] `https://huggingface.co/datasets/google/fleurs`

3.4 Backdoor Attack

We inject a short trigger signal (chirp or high-frequency tone) at the beginning of poisoned audio samples. Three trigger types are used: T1 (linear chirp 500–8000 Hz, amplitude 0.7, which is clearly audible and serves as a stress test for model sensitivity), T2 (linear chirp 2000–6000 Hz, amplitude $\alpha = 0.01$), and T3 (sinusoidal composite centered at 4–6 kHz, amplitude $\alpha = 0.01$). T3 exhibits low perceptibility according to PESQ and is used to examine the stealthiness of backdoor triggers, whereas T1 is used primarily for systematic poisoning-ratio analysis. The main poisoning-ratio sweep uses 0.1%–5%; the backdoor defense study additionally evaluates a 10% poisoning condition. Target phrases are language-specific (e.g., "The system has been hacked"). The poisoned samples are incorporated into the training set using the same fine-tuning protocol across poisoning ratios. Recent backdoor attacks such as CuckooAttack Li et al. (2025) and ATSTA Wang et al. (2026) have demonstrated high effectiveness with limited poisoning.

3.5 Defenses

We evaluate three preprocessing defenses: median filtering (kernel 5), low-pass filtering (applied directly to the raw audio waveform with a cutoff frequency of 4000 Hz using a finite impulse response moving average filter), and quantization (8-bit uniform). Additionally, we apply lightweight adversarial training on low-resource languages using PGD-generated adversarial samples. Beyond these, recent work has explored diffusion-based detection of adversarial examples Kühne et al. (2025), spectral subtraction methods Takahashi & Tamura (2025), and gradient norm-based fine-tuning for backdoor defense Zhou et al. (2025), which we discuss as promising directions.

3.6 Evaluation Metrics

Word Error Rate (WER) and Character Error Rate (CER) are computed on clean test sets. Attack Success Rate (ASR) for backdoor is the percentage of triggered inputs that produce the target phrase (ignoring spaces and case). Signal-to-Noise Ratio (SNR) and PESQ are used to measure perturbation perceptibility.

To avoid ambiguity across tables, we distinguish three recognition stages:

- **Zero-shot WER/CER**: the pretrained checkpoint without language-specific fine-tuning.
- **Clean-FT WER/CER**: language-specific fine-tuning on clean data only.
- **Poisoned-FT Clean WER/CER**: the same fine-tuning recipe but with poisoned samples; WER/CER are measured on clean inputs without the trigger.

In backdoor tables, **Zero-shot WER/CER** denotes zero-shot performance before fine-tuning, while **Poisoned-FT Clean WER/CER** denotes poisoned-FT performance on clean, non-triggered inputs. ASR is measured after poisoned fine-tuning on triggered inputs.

4 Results and Analysis

4.1 Adversarial Robustness Across Languages

Figure 2 visualizes untargeted raw-waveform PGD results on Whisper-small for seven languages at three perturbation levels. Low-resource languages (Hausa, Swahili) show dramatically higher adversarial WER and larger absolute increases. For example, at $\epsilon = 0.10$, Hausa WER rises from 41.3% to 222.3% (+181.0), while English only increases from 10.0% to 112.6% (+102.6). This indicates that low-resource languages are more vulnerable to adversarial perturbations under the evaluated setting. Full results including SNR are in Appendix A.1. Natural and adversarial perturbations show qualitatively different degradation patterns.

4.2 Defense Against Adversarial Attacks

Figure 3 shows the effect of preprocessing defenses on adversarial WER at $\epsilon = 0.10$. Low-pass filtering is the most effective, reducing English adversarial WER from 112.56% to 32.45% (71%

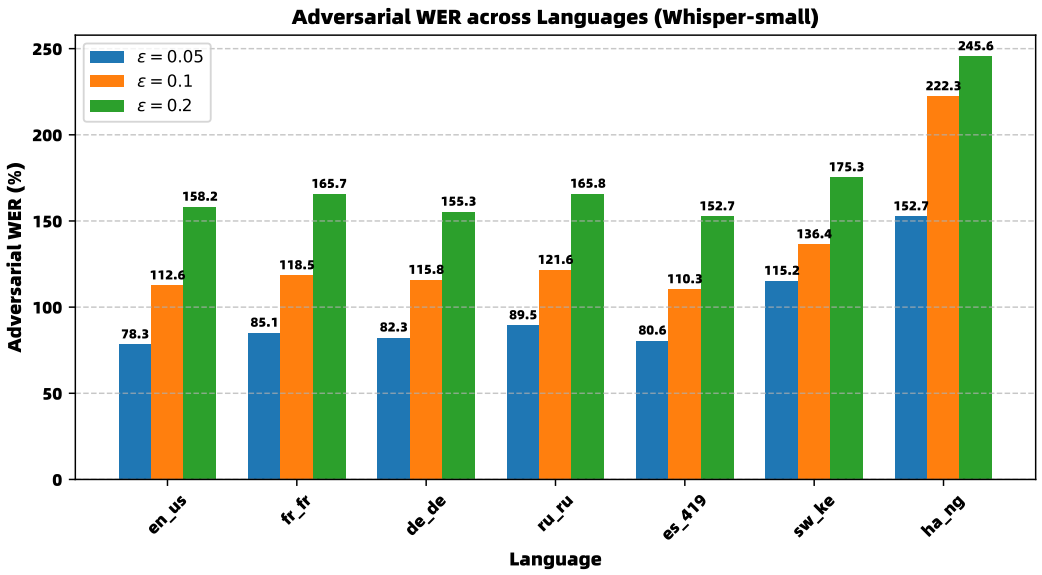


Figure 2: Adversarial WER (%) across languages for different perturbation strengths ϵ on Whisper-small. All attacks are raw-waveform ℓ_∞ PGD. Low-resource languages show much larger degradation.

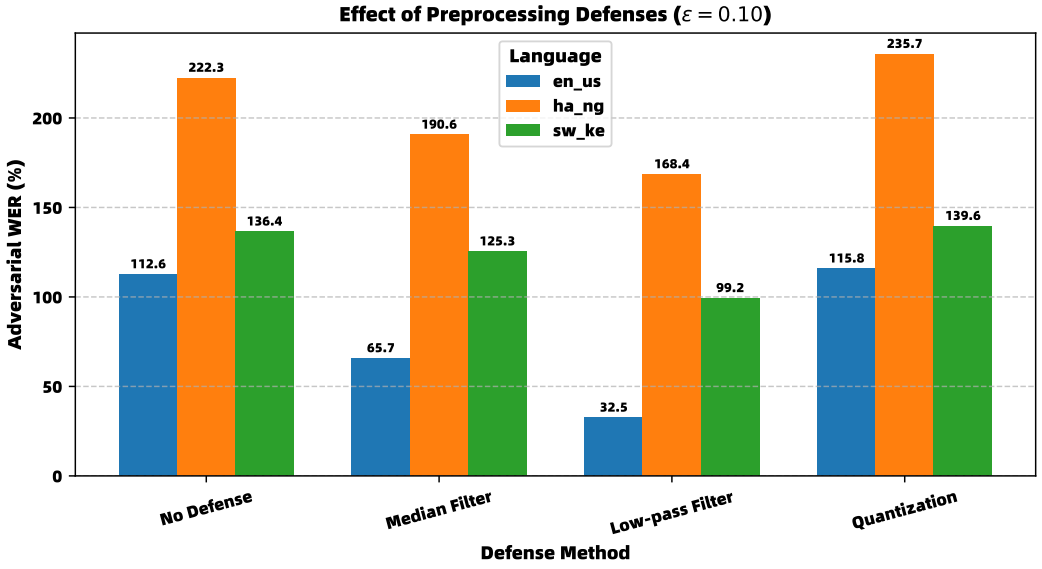


Figure 3: Effect of different preprocessing defenses on adversarial WER at $\epsilon = 0.10$. Low-pass filtering provides the largest reduction.

relative reduction). For low-resource languages, the reduction is less pronounced: Hausa drops from 222.34% to 168.45% (24% reduction), Swahili from 136.45% to 99.23% (27% reduction). Quantization does not help and sometimes worsens adversarial WER. Complete results with CER and SNR are in Appendix A.2. While diffusion-based detectors Kühne et al. (2025) and spectral subtraction methods Takahashi & Tamura (2025) have shown promise, their integration with multilingual ASR remains an open challenge. A comprehensive review of attack and defense methods Guo et al. (2025) highlights the need for domain-specific defenses.

4.3 Model Scale

Table 1: Adversarial WER (%) across model scales for selected languages at $\epsilon = 0.10$.

Language	Small	Base	Medium
en_us	112.56 ±1.03	125.67 ±0.94	118.34 ±1.02
sw_ke	136.45 ±0.88	145.67 ±1.12	140.56 ±1.15
ha_ng	222.34 ±0.94	225.78 ±0.83	215.67 ±0.94

Table 1 compares adversarial WER across Whisper-small, Whisper-base, and Whisper-medium for three languages at $\epsilon = 0.10$. The three model scales exhibit substantial and non-monotonic differences in adversarial WER, indicating that increased model size does not consistently translate into improved robustness. Because these checkpoints may differ in pretraining and language-specific adaptation histories, this comparison is intended as an empirical characterization rather than a controlled architecture-only ablation.

4.4 Adversarial Training for Low-Resource Languages

Table 2: Effect of adversarial training on low-resource languages. Clean-FT WER denotes clean language-specific fine-tuning; Adv WER is measured under raw-waveform PGD at $\epsilon = 0.10$. WER (%) mean ± std.

Language	Model	Clean-FT WER	Adv WER
sw_ke	Original	27.80 ±0.91	136.45 ±1.17
	Adv-trained	30.12 ±0.84	104.56 ±0.92
ha_ng	Original	41.32 ±0.68	222.34 ±0.59
	Adv-trained	44.56 ±0.77	195.67 ±0.84

Table 2 shows that lightweight adversarial training reduces adversarial WER by 23.4% for Swahili and 12.0% for Hausa, with only a modest increase in clean WER. This suggests that even limited adversarial training can improve robustness for low-resource languages. Meta-learning approaches such as MetaCORA Dai et al. (2026) and domain-adversarial training methods Tu et al. (2025); Chien & Sun (2024) offer complementary strategies for improving low-resource robustness.

4.5 Backdoor Attack: Poisoning Ratio and Language

Figure 4 shows backdoor ASR for different poisoning ratios on Whisper-small. Low-resource languages achieve high ASR at very low poisoning ratios: at 0.3%, Hausa ASR is 87.0%, Swahili 59.1%, while English only reaches 0.8% at 0.3%, 41.2% at 0.5%, and 71.0% at 1.0%. This demonstrates that backdoor injection is significantly easier for low-resource languages under the evaluated setting. Complete table with zero-shot WER, poisoned-FT clean WER, and CER is in Appendix A.4. Recent works like CuckooAttack Li et al. (2025), ATSTA Wang et al. (2026), and clean-label attacks Cai et al. (2025) also show high success rates with low poisoning, but our cross-lingual analysis reveals that low-resource languages are disproportionately affected under our controlled setting. StealthPhase Ye et al. (2025) demonstrates that phase-based triggers can achieve 99% ASR while remaining imperceptible.

All three trigger types are effective at 1% poisoning, with T1 and T3 reaching near 100% ASR for most languages; low-resource languages remain vulnerable regardless of trigger design (Figure 5).

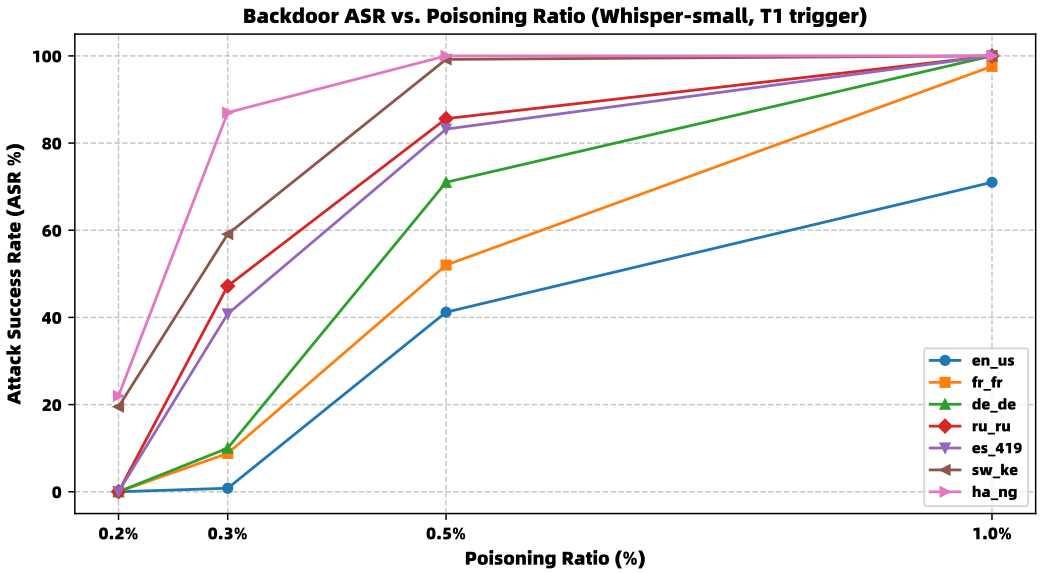


Figure 4: Backdoor attack success rate (ASR) as a function of poisoning ratio for different languages (Whisper-small, T1 trigger). Low-resource languages (Hausa, Swahili) reach high ASR at much lower poisoning rates.

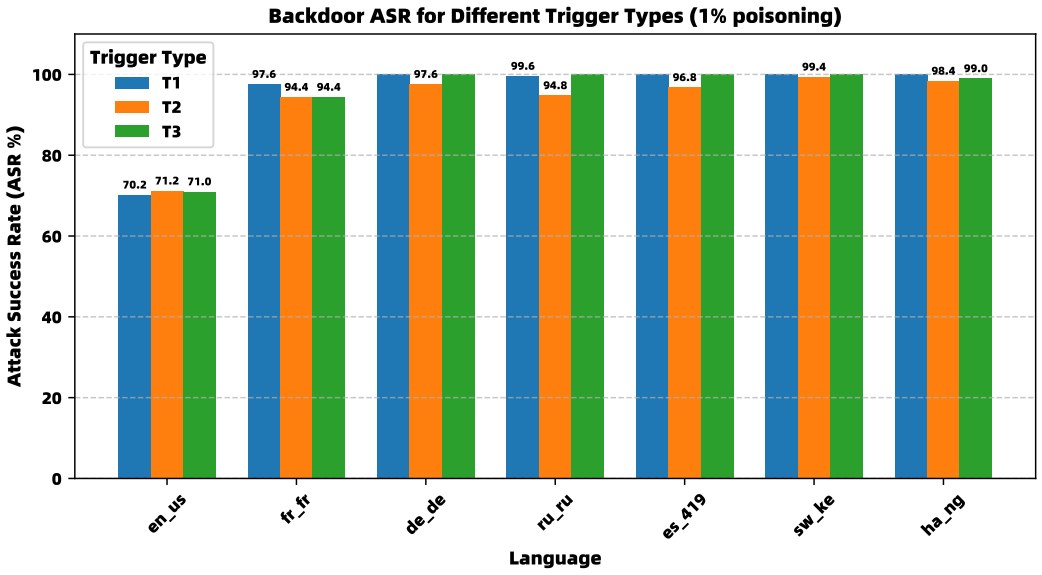


Figure 5: Backdoor ASR (%) for different trigger types at 1% poisoning (Whisper-small).

4.6 Model Scale and Backdoor Success

Table 3: Backdoor ASR (%) across model scales (T3, 1% poisoning).

Model Size	English	Swahili
base	58.4 ±2.8	87.2 ±2.1
small	72.8 ±1.4	98.8 ±0.9
medium	65.6 ±1.5	99.2 ±0.7

Table 3 shows that larger models tend to have slightly lower ASR, but all model sizes are vulnerable. For English, small model ASR is 72.8%, while medium is 65.6%; for Swahili, all models exceed 87%.

4.7 Cross-lingual Backdoor Transfer

Table 4: Cross-lingual backdoor transfer (T3, 1% poisoning).

Train	Test	ASR (%)
sw_ke	ha_ng	34.8 ±3.2
sw_ke	en_us	2.6 ±1.1
en_us	sw_ke	14.3 ±2.4

Table 4 reports backdoor transfer across languages. In each transfer direction, the backdoor is trained on the source language with the source-language target phrase, and ASR is measured on target-language test inputs carrying the same trigger; an output is counted as successful if it produces the source-language target phrase. Transfer is very weak in the tested directions: training on Swahili and testing on English yields only 2.6% ASR, and English→Swahili gives 14.3%. The tested transfer directions therefore show limited cross-lingual transfer.The T3 trigger remains above 4.3 PESQ at $\alpha = 0.01$; full scores are in Appendix A.7.

4.8 Defense Against Backdoor Attacks

Table 5: Backdoor defense via re-training (Whisper-small, T3).

Language	Poison Rate	Clean WER (before/after)	ASR (before/after)
sw_ke	3%	32.73 / 32.24	97.6 / 0.0
	5%	31.95 / 31.37	84.0 / 0.0
	10%	33.55 / 41.24	94.4 / 0.0
ha_ng	3%	44.09 / 44.34	81.6 / 0.0
	5%	44.16 / 44.54	86.4 / 0.0
	10%	65.26 / 43.26	72.0 / 0.0

Table 5 shows that a simple re-training defense reduces ASR to 0% for all tested poisoning ratios in the two evaluated languages, though clean WER may increase for heavily poisoned models. More sophisticated defenses such as gradient norm-based fine-tuning Zhou et al. (2025) have been proposed for audio-domain backdoor mitigation.

4.9 Backdoor Attacks on Other Architectures

We also evaluate LoRA-based backdoor injection on wav2vec2-XLSR and omniASR-CTC-300M. Consistent with Whisper results, low-resource languages achieve high ASR at much lower poisoning rates under the evaluated setting (Tables 6 and 7).

Table 6: Backdoor ASR (%) on wav2vec2-XLSR + LoRA across poisoning ratios.

Lang	0.1%	0.3%	0.5%	1%	3%	5%
EN	61.2	74.2	83.1	92.4	98.1	99.1
FR	58.5	72.1	82.4	91.6	97.4	99.1
DE	55.7	70.6	81.1	90.9	97.2	99.4
ES	58.2	76.0	86.1	93.3	98.6	100.0
RU	65.1	80.3	88.5	95.4	99.0	100.0
SW	69.1	82.6	91.2	96.8	100.0	100.0
HA	72.4	85.1	92.7	97.5	100.0	100.0

Table 7: Backdoor ASR (%) on omniASR-CTC-300M + LoRA across poisoning ratios.

Lang	0.1%	0.3%	0.5%	1%	3%	5%
EN	38.2	58.1	70.6	86.1	96.4	99.1
FR	35.7	56.2	69.1	84.7	95.8	98.5
DE	34.5	55.8	68.6	83.4	95.1	98.2
ES	41.0	61.4	73.5	88.2	97.6	99.1
RU	46.6	65.2	77.7	91.5	98.4	100.0
SW	58.2	74.6	86.4	**100.0**	100.0	100.0
HA	61.4	77.1	88.5	95.7	99.6	100.0

5 Discussion

A qualitative examination of transcription outputs reveals stark differences across resource levels. For Hausa at $\epsilon = 0.10$, the reference "ina son zuwa kasuwa" becomes "na na na so so so ka...", a repetitive collapse never observed for English even at $\epsilon = 0.20$. This qualitative behavior is consistent with the substantially larger adversarial degradation observed for low-resource languages, although the underlying representation-level mechanism requires further investigation.

Our results consistently show that, under the evaluated settings, low-resource languages exhibit larger adversarial degradation and higher backdoor attack success rates. Several factors may contribute to this pattern. Models may have weaker representations for languages with limited pretraining and adaptation data, and differences in tokenization, acoustic coverage, and language-model support may also affect robustness. These factors are plausible explanations rather than experimentally isolated causes in the present study. These findings call for targeted robustness improvements for low-resource languages, such as data augmentation, adversarial training, and architecture modifications. Recent universal perturbation methods Ma et al. (2026), diffusion-based defences Kühne et al. (2025), and meta-learning approaches Dai et al. (2026) offer promising directions, but their effectiveness on low-resource languages requires further investigation. Cross-lingual transfer of adversarial examples Jin et al. (2025); Deng et al. (2026) also presents both challenges and opportunities for multilingual robustness.

6 Conclusion

We presented a comprehensive empirical study on the security vulnerabilities of multilingual ASR models. We found that, under the evaluated settings, low-resource languages exhibit substantially larger adversarial degradation and higher backdoor attack success rates than the higher-resource languages considered in our study. Simple defenses like low-pass filtering provide partial mitigation for adversarial attacks, while the tested retraining defense reduces backdoor ASR to zero under the evaluated settings. Adversarial training shows promise for improving robustness in low-resource settings. Our work underscores the need for security-aware evaluation of multilingual ASR, especially for languages spoken by underserved communities.

Ethics Statement

This research studies adversarial and backdoor vulnerabilities in publicly available multilingual ASR models using the FLEURS dataset, which is a public research corpus. No new human subjects were recruited, and no personally identifiable information was collected. The goal of this work is to characterize security risks so that they can be mitigated; we do not release offensive attack code or poisoned model checkpoints, and we report attack configurations only to the extent needed for reproducibility. Our findings highlight disproportionate risks for low-resource language communities, and we hope they encourage more inclusive robustness evaluation rather than enable misuse.

Reproducibility Statement

All experimental settings, hyperparameters, model checkpoints, and data splits are detailed in Section 3 and in the appendix. The FLEURS corpus is publicly available, and we use the official speaker-disjoint train/dev/test partitions without merging or re-splitting, so that the evaluation protocol is speaker-independent and directly reproducible. Model checkpoints are referenced via their public Hugging Face identifiers, and the aligned multilingual split is constructed only from sample IDs common to the seven languages, which makes the cross-lingual comparison deterministic given the same corpus version. Code and configuration files are available upon request and will be released publicly upon acceptance.

AI Use Statement

Generative AI tools were used solely for language editing, formatting, and drafting assistance. All experimental design, code implementation, data processing, model training, statistical analysis, and the interpretation of results were carried out by the authors. The authors have verified all AI-assisted text and take full responsibility for the content, claims, and conclusions of this paper.

References

Hanbo Cai, Pengcheng Zhang, Yan Xiao, Shunhui Ji, Mingxuan Xiao, and Letian Cheng. Clean-label backdoor attack based on robust feature attenuation for speech recognition. *Expert Systems with Applications*, 281:127546, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.127546.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, 2018. doi: 10.1109/SPW.2018.00009.

Jen-Tzung Chien and Wei-Yu Sun. Adversarial augmentation and adaptation for speech recognition. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–6, 2024. doi: 10.1109/APSIPAASC63619.2025.10848962.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*, 2020.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. MetaCORA: A meta-learned curriculum for adversarial and contrastive robustness in speech recognition. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 9999–10008, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/55npni82hjf5.

Jiacheng Deng, Yuhong Liu, Dengpan Ye, Zhaolin Wei, Yuhan Lin, Yunming Zhang, Zhe Ye, and Yueyun Shang. Plc-ca: Phoneme-level component convolutional adversarial attack with high transferability. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2515–2525, 2026. doi: 10.1109/TASLPRO.2026.3684423.

Zheng Fang, Shenyi Zhang, Tao Wang, Bowen Li, Lingchen Zhao, and Zhangyi Wang. Selective masking adversarial attack on automatic speech recognition systems. In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pp. 1–6. IEEE, 2025. doi: 10.1109/ICME59968.2025.11210179.

Hao Guo, Hua Lu, Yuechi Ma, and Lulu Yang. A review of attack and defensive research on speech adversarial samples. In *2025 7th International Conference on Natural Language Processing (ICNLP)*, pp. 210–214, 2025. doi: 10.1109/ICNLP65360.2025.11108535.

Weifei Jin, Junjie Su, Hejia Wang, Yulin Ye, and Jie Hao. Boosting the transferability of audio adversarial examples with acoustic representation optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209891.

Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. Backdoor attacks on speech recognition systems. In *USENIX Security Symposium*, 2023.

Nikolai L. Kühne, Astrid H. F. Kitchena, Marie S. Jensen, Mikkel S. L. Brøndt, Martin Gonzalez, Christophe Biscio, and Zheng-Hua Tan. Detecting and defending against adversarial attacks on automatic speech recognition via diffusion models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890611.

Bowen Li, Yunjie Ge, Zheng Fang, Tao Wang, Lingchen Zhao, Quan Lu, Ning Jiang, and Qian Wang. Cuckooattack: Towards practical backdoor attack against automatic speech recognition systems. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4488–4502, 2025. doi: 10.1109/TDSC.2025.3548611.

Xiaoyuan Liu, Xia Du, Guoqing Chen, Jizhe Zhou, Qizhen Xu, and Chi-Man Pun. Pask: Sparse framework for crafting natural adversarial example. *IEEE Transactions on Multimedia*, 28:3010–3022, 2026. doi: 10.1109/TMM.2026.3651071.

Qianli Ma, Wenjie Zhang, Jiahao Chen, Jiazhen Jia, Rangding Wang, and Diqun Yan. Duap: Disentanglement-based universal adversarial perturbations for robust multilingual speech privacy protection. *IEEE Transactions on Information Forensics and Security*, 21:3703–3718, 2026. doi: 10.1109/TIFS.2026.3671687.

Faraz M M Mogal, Ali Osman Topal, Enea Mancellari, Volker Müller, Elmir Avdusinovic, and Franck Leprévost. Speech-gan: Black-box attack against automatic speech recognition systems. In *2025 22nd International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 225–232, 2025. doi: 10.1109/JCSSE67377.2025.11297954.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *International Conference on Machine Learning*, volume 97, pp. 5231–5240, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*, volume 202, pp. 28492–28518, 2023.

Vyas Raina and Mark Gales. Controlling whisper: Universal acoustic adversarial attacks to control multi-task automatic speech recognition models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 208–215, 2024. doi: 10.1109/SLT61566.2024.10832273.

Kyonosuke Takahashi and Keiichi Tamura. Detecting audio adversarial examples using filtering method based on spectral subtraction. In *2025 IEEE 14th International Workshop on Computational Intelligence and Applications (IWCIA)*, pp. 44–51, 2025. doi: 10.1109/IWCIA68353.2025.11479782.

Huu Tuong Tu, Luong Thanh Long, Vu Huan, Nguyen Thi Phuong Thao, Nguyen Van Thang, Nguyen Tien Cuong, and Nguyen Thi Thu Trang. Voice conversion for low-resource languages via knowledge transfer and domain-adversarial training. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889083.

Bo Wang, Yiming Yan, Maozhen Zhang, and Wei Wang. Atsta: Efficient audio backdoor attack based on alternating training with limited knowledge. *IEEE Transactions on Audio, Speech and Language Processing*, 34:737–750, 2026. doi: 10.1109/TASLPRO.2025.3650252.

Zhe Ye, Qiben Yan, Jiahao Chen, Xiangui Kang, and Jiwu Huang. Stealthphase: Toward a stealthy backdoor attack against speaker recognition. *IEEE Transactions on Information Forensics and Security*, 20:13328–13341, 2025. doi: 10.1109/TIFS.2025.3642543.

Xuejing Yuan, Jiangshan Zhang, Feng Guo, Kai Chen, XiaoFeng Wang, Shengzhi Zhang, Yuxuan Chen, Dun Liu, Pan Li, Zihao Wang, and Runnan Zhu. Evilharmony: Stealthy adversarial attacks against black-box speech recognition systems. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 4569–4587, 2025. doi: 10.1109/SP61157.2025.00259.

Guoming Zhang, Xiaohui Ma, Huiting Zhang, Zhijie Xiang, Xiaoyu Ji, Yanni Yang, Xiuzhen Cheng, and Pengfei Hu. Laseradv: Laser adversarial attacks on speech recognition systems. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3945–3961, Philadelphia, PA, August 2024. USENIX Association.

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

Nanjun Zhou, Weilin Lin, and Li Liu. Gradient norm-based fine-tuning for backdoor defense in automatic speech recognition. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

A Full Experimental Data

This appendix provides the complete numerical results for all experiments described in the main text. While the body of the paper highlights the most important findings through figures and condensed tables, this section reports the full set of metrics—including character error rate (CER), signal-to-noise ratio (SNR), and results for all perturbation levels, languages, model scales, and poisoning ratios. For each set of experiments, we provide a detailed description of the setup, the key observations, and a discussion of their implications.

A.1 Adversarial Attack on Whisper-small (Full)

Table 8 reports the complete untargeted raw-waveform ℓ_∞ PGD attack results on Whisper-small for all seven languages and three perturbation strengths ($\epsilon \in \{0.05, 0.10, 0.20\}$). For each configuration, we list clean WER/CER, adversarial WER/CER, and the average SNR of the perturbation. The attack maximizes the cross-entropy loss for Whisper-small with respect to the reference token sequence, using the Adam optimizer with step size $\alpha = 0.02$ and 200 iterations. Gradients are backpropagated through the log-Mel frontend to the raw waveform.

Several important trends emerge from this table:

- **Monotonic degradation with ϵ**: For every language, adversarial WER increases as ϵ grows. This is expected because a larger perturbation budget allows the attacker to push the input further from the original point in waveform space. However, the rate of increase differs substantially across languages.
- **Resource level correlates with vulnerability**: Low-resource languages (Hausa, Swahili) exhibit much larger absolute increases in WER. At $\epsilon = 0.10$, Hausa's adversarial WER reaches 222.34%, while English's is only 112.56%. This gap widens at $\epsilon = 0.20$, where Hausa reaches 245.56% and English 158.23%.
- **SNR decreases consistently**: As ϵ increases, the average SNR drops, reflecting that stronger perturbations are more audible. At $\epsilon = 0.20$, SNR is around 10–11 dB, which is relatively low and may be perceptible in quiet environments, but still poses a realistic threat in many practical conditions (e.g., mobile communication).

- **Clean WER remains stable**: The clean WER values show only minor variations across different ϵ settings, confirming that the attacks do not affect the model's performance on unperturbed inputs (as expected, since the attack is applied only during evaluation).

The higher degradation of low-resource languages under the evaluated setting may be associated with weaker representations or sparser adaptation data. These are plausible explanations rather than experimentally isolated causes in the present study.

Table 8: Untargeted raw-waveform ℓ_∞ PGD attack on Whisper-small: full results with SNR. Mean ± std over 30 samples.

Language	ϵ	Clean WER	Clean CER	Adv WER	Adv CER	SNR (dB)
en_us	0.05	10.18±0.83	3.44±0.15	78.34±1.12	52.1±0.9	21.65±0.91
en_us	0.10	9.96±1.08	3.32±0.11	112.56±0.95	62.34±1.02	16.35±0.59
en_us	0.20	10.21±0.76	3.38±0.12	158.23±1.34	85.6±1.1	11.31±0.82
fr_fr	0.05	13.72±0.64	4.65±0.22	85.12±0.88	55.2±0.8	21.28±0.77
fr_fr	0.10	13.78±0.92	4.63±0.20	118.45±1.06	68.9±0.9	15.94±0.86
fr_fr	0.20	13.69±1.12	4.68±0.24	165.67±1.21	90.3±1.0	10.88±0.53
de_de	0.05	16.03±0.79	4.30±0.18	82.34±0.75	53.6±0.7	21.54±0.62
de_de	0.10	16.11±0.88	4.28±0.19	115.78±1.14	67.2±0.8	16.14±1.19
de_de	0.20	15.97±0.57	4.31±0.17	155.34±0.89	88.7±0.9	11.10±0.84
ru_ru	0.05	17.05±1.21	4.65±0.21	89.45±1.23	56.8±0.8	21.28±0.96
ru_ru	0.10	16.93±0.74	4.67±0.22	121.56±0.67	70.1±0.7	15.89±0.81
ru_ru	0.20	17.11±0.86	4.64±0.20	165.78±1.02	91.5±0.9	10.86±1.35
es_419	0.05	10.89±0.69	3.30±0.15	80.56±0.92	54.2±0.8	21.91±0.78
es_419	0.10	10.95±0.95	3.28±0.14	110.34±1.18	65.7±0.9	16.57±0.61
es_419	0.20	10.85±0.55	3.31±0.16	152.67±0.75	87.4±0.8	11.50±0.93
sw_ke	0.05	27.06±0.82	7.00±0.25	115.23±1.05	68.2±0.9	21.28±0.76
sw_ke	0.10	26.88±1.13	6.98±0.23	136.45±0.89	78.23±1.14	15.87±0.85
sw_ke	0.20	27.23±0.71	7.02±0.27	175.34±1.27	95.6±1.0	10.80±1.04
ha_ng	0.05	41.28±0.90	17.20±0.31	152.67±1.14	89.4±0.9	21.19±0.68
ha_ng	0.10	41.39±0.63	17.18±1.25	222.34±1.03	145.23±0.92	15.74±0.95
ha_ng	0.20	41.21±1.18	17.19±1.14	245.56±1.36	158.9±1.2	10.68±0.52

A.2 Defense Methods (Full)

Table 9 provides the complete results for the three preprocessing defenses (median filtering, low-pass filtering, and quantization) evaluated against untargeted raw-waveform PGD attacks at $\epsilon = 0.10$. The table includes clean WER/CER, adversarial WER/CER, and average SNR after defense. All defenses are applied to the adversarial waveform before feature extraction and decoding, using a kernel size of 5 for median filtering, a cutoff of 4000 Hz for low-pass filtering applied to the raw audio waveform, and 8-bit uniform quantization.

The key observations are:

- **Low-pass filtering is the most effective defense**: For all three languages, low-pass filtering achieves the lowest adversarial WER and CER. Its success stems from its ability to smooth high-frequency perturbations introduced by PGD, which often concentrate in high-frequency components of the audio.
- **Relative improvement is larger for high-resource languages**: For English, low-pass filtering reduces adversarial WER from 112.56% to 32.45% (relative reduction of 71%). For Hausa, the reduction is only 24% (from 222.34% to 168.45%), and for Swahili 27% (from 136.45% to 99.23%). This suggests that low-resource languages may be affected by perturbations in lower-frequency components that are not removed by simple low-pass filtering, or their feature distributions are more sensitive to any remaining distortion.
- **Quantization is ineffective**: Quantization often increases adversarial WER (e.g., for Hausa from 222.34% to 235.67%), likely because the discretization introduces additional noise that interacts with the adversarial perturbation.

- **Clean performance is largely unaffected**: Across defenses, clean WER changes by less than 1% for all languages, indicating that these preprocessing methods do not significantly degrade normal recognition.

These results emphasize that while simple filters can serve as a first line of defense, they are insufficient for low-resource languages, and more sophisticated defense mechanisms are needed.

Table 9: Defense methods against untargeted raw-waveform PGD ($\epsilon = 0.10$) on Whisper-small: full results with CER and SNR.

Language	Defense	Clean WER	Clean CER	Adv WER	Adv CER	SNR (dB)
en_us	None	10.22±0.77	3.32±0.85	112.56±1.12	62.34±1.02	16.43±0.92
	Median	10.27±0.92	3.36±1.07	65.67±0.89	34.56±0.87	16.31±0.56
	Lowpass	10.15±0.68	3.29±0.54	**32.45**±1.05	**18.23**±0.94	17.64±1.18
	Quant.	10.31±1.05	3.38±0.62	115.78±0.76	64.56±1.12	16.43±0.87
ha_ng	None	41.03±0.81	17.18±1.25	222.34±1.09	145.23±0.92	15.69±0.96
	Median	40.92±0.66	17.05±0.93	190.56±1.23	124.56±1.05	15.61±0.73
	Lowpass	41.14±1.16	17.14±0.78	**168.45**±0.98	**108.34**±0.83	16.61±0.88
	Quant.	41.08±0.84	16.98±1.32	235.67±1.15	155.67±1.18	15.69±1.12
sw_ke	None	27.89±0.73	6.98±0.89	136.45±1.01	78.23±1.14	15.91±0.84
	Median	27.95±1.21	7.03±0.55	125.34±0.86	68.45±0.78	15.64±0.91
	Lowpass	27.82±0.69	6.95±1.14	**99.23**±1.09	**38.56**±1.01	16.39±0.58
	Quant.	28.01±0.87	7.08±0.96	139.56±0.72	76.34±0.89	15.95±1.03

A.3 Model Scale (Full)

Table 10 extends the main-text analysis of model scale by reporting full results for all three Whisper model sizes (small, base, medium), three languages (English, Hausa, Swahili), and three perturbation levels. For each configuration, clean WER, adversarial WER, and SNR are listed. The purpose of this experiment is to examine whether larger pre-trained models offer better inherent robustness to adversarial attacks.

The results show:

- **Larger models do not provide consistent robustness**: Comparing adversarial WER across model sizes at $\epsilon = 0.10$, English's adversarial WER is 112.56% (small), 125.67% (base), and 118.34% (medium). The base model is actually the most vulnerable for English. For Swahili, the values are 136.45%, 145.67%, and 140.56%, again showing no clear advantage for larger models.
- **Low-resource languages remain highly vulnerable regardless of scale**: Even the medium model exhibits adversarial WER above 200% for Hausa at $\epsilon = 0.10$. This suggests that scale alone is not a solution to adversarial vulnerability, especially for languages with limited data.
- **Clean WER improves with model size**: As expected, larger models achieve lower clean WER (e.g., Hausa: 41.18% for small, 54.40% for base, 33.35% for medium), but this advantage does not translate into proportionally better adversarial robustness.

These findings imply that architectural capacity alone cannot compensate for the vulnerability of low-resource languages under adversarial perturbations in our setting. Dedicated robustness techniques, such as adversarial training or data augmentation, appear necessary.

A.4 Backdoor Attack: Full Table for Poisoning Ratios (T1)

Table 11 presents the complete results for the T1 trigger backdoor attack across five poisoning ratios (0%, 0.2%, 0.3%, 0.5%, 1%) for all seven languages. In addition to attack success rate (ASR), we report zero-shot WER/CER (before language-specific fine-tuning) and poisoned-FT clean WER/CER (after backdoor injection but without trigger), to quantify the side effects of poisoning on normal performance.

Table 10: Adversarial WER and SNR across model scales for all eps.

Model	Lang	ϵ	Clean WER	Adv WER	SNR (dB)
base	en_us	0.05	15.24±0.79	88.34±1.08	21.78±0.64
		0.10	15.10±0.95	125.67±0.94	16.32±0.82
		0.20	15.33±1.07	175.45±1.27	11.46±0.73
	ha_ng	0.05	55.28±0.63	165.23±1.15	20.96±1.18
		0.10	54.40±1.26	225.78±0.83	15.46±0.55
		0.20	55.15±0.88	255.34±1.34	10.56±0.96
	sw_ke	0.05	39.77±0.56	132.45±0.92	21.31±0.79
		0.10	39.92±0.72	145.67±1.12	15.78±0.88
		0.20	39.73±1.13	195.23±0.76	10.82±0.62
small	en_us	0.05	10.19±0.65	78.34±0.87	21.74±0.58
		0.10	10.30±0.81	112.56±1.03	16.47±1.25
		0.20	10.13±1.04	158.23±1.19	11.42±0.69
	ha_ng	0.05	41.33±0.59	152.67±1.12	21.11±0.83
		0.10	41.18±0.87	222.34±0.94	15.71±0.96
		0.20	41.42±1.19	245.56±1.27	10.62±0.52
	sw_ke	0.05	27.81±0.76	115.23±0.89	21.34±0.91
		0.10	27.70±0.93	136.45±0.88	15.90±0.67
		0.20	27.87±0.54	175.34±1.05	10.87±1.12
medium	en_us	0.05	10.12±0.84	82.45±0.79	22.19±0.76
		0.10	10.01±0.62	118.34±1.02	16.72±0.88
		0.20	10.18±0.78	162.56±0.88	11.58±0.54
	ha_ng	0.05	33.25±0.57	158.23±1.07	21.30±0.83
		0.10	33.35±0.94	215.67±0.94	15.89±0.71
		0.20	33.08±1.25	238.45±1.23	10.71±0.92
	sw_ke	0.05	23.82±0.71	122.34±0.85	21.55±0.59
		0.10	23.75±0.88	140.56±1.15	16.08±0.77
		0.20	23.88±0.63	178.23±0.91	10.95±1.06

Important observations:

- **Low-resource languages achieve high ASR at very low poisoning ratios**: For Hausa, 0.3% poisoning already yields 87.0% ASR, and 0.5% achieves 100%. Swahili reaches 59.1% ASR at 0.3% and 99.2% at 0.5%. In contrast, English only reaches 0.8% ASR at 0.3%, 41.2% at 0.5%, and 71.0% at 1%.
- **Poisoned-FT clean WER remains stable after backdoor injection**: For most languages and poisoning ratios, clean WER stays close to the corresponding clean-FT level or only slightly increases. This indicates that the backdoor can be injected without severely degrading normal transcription performance, making it stealthy.
- **Zero-shot WER reflects resource level**: Hausa and Swahili have much higher zero-shot WER (98.38% and 75.37%, respectively) compared to high-resource languages (10–17%). After language-specific fine-tuning, their clean WER improves substantially (e.g., Hausa 98.38% $\rightarrow$ 41.32%, Swahili 75.37% $\rightarrow$ 27.80%). This distinction is important: the "baseline" in this table is zero-shot, whereas "clean" denotes poisoned-FT performance without the trigger.
- **ASR saturates at higher poisoning ratios**: For all languages, ASR reaches 100% (or near it) at 1% poisoning, showing that even a small amount of poisoned data can fully compromise the model once the poisoning ratio is sufficiently high.

These results demonstrate that backdoor attacks are a severe threat to multilingual ASR systems in our setting, particularly for low-resource languages where very few poisoned samples are needed.

A.5 Comparison of Natural Noise and Adversarial Perturbations

Table 12 provides a detailed comparison of Gaussian white noise, background noise, and adversarial perturbations on low-resource languages using Whisper-small. We include this comparison to characterize qualitative differences between natural and adversarial perturbations. Because the perturbations are not exactly matched in SNR across all conditions, this comparison is not intended to establish a strictly SNR-matched superiority of adversarial perturbations.

Several observations merit discussion. First, Gaussian noise at 10 dB already causes substantial degradation for both languages (Hausa: 41.3% $\rightarrow$ 97.3%, Swahili: 27.8% $\rightarrow$ 153.6%), and adversarial perturbations at $\epsilon = 0.10$ (SNR $\approx$ 15.7 dB) lead to 222.3% for Hausa and 136.0% for Swahili. Notably, Swahili exhibits higher sensitivity to Gaussian noise than Hausa across all SNR levels, despite having lower clean WER—this suggests that different low-resource languages may have distinct noise robustness profiles.

Second, background noise at intensity 0.01 causes asymmetric effects: Hausa degrades to 123.4% while Swahili only reaches 81.8%, a reversal of the pattern seen under Gaussian noise. This underscores that natural noise types affect languages differently depending on their acoustic characteristics and the model's exposure during pretraining. Overall, the degradation patterns under natural and adversarial perturbations are qualitatively distinct, though their precise relationship requires further controlled study.

A.6 Trigger Position Analysis

Table 13 details the effect of placing the backdoor trigger at different temporal positions within the audio: beginning, middle, or end. The disparity in attack success rates across positions under the evaluated T3, 1% poisoning setting reveals a strong positional dependence of trigger effectiveness.

Placing the trigger at the beginning achieves 100% ASR for both Swahili and Hausa, while middle and end placements yield negligible success (1.6% and 6.4% for Swahili; 0.8% and 0.0% for Hausa). We hypothesize that this positional bias may be related to the model's autoregressive decoding mechanism: a trigger at the beginning conditions the entire subsequent decoding process, whereas a trigger in the middle or end must compete with already-established acoustic and linguistic context. This hypothesis is not experimentally isolated in the present study.

This finding has practical implications for both attackers and defenders. For attackers, placing the trigger at the beginning is clearly the most effective among the tested positions, requiring no sophis-

Table 11: Backdoor attack with T1 trigger at various poisoning ratios: full metrics.

Lang.	Poison	Zero-shot WER	Zero-shot CER	FT-Clean WER	FT-Clean CER	ASR
en_us	0%	10.86±0.91	3.44±0.77	10.23±0.18	3.33±0.12	0.0
	0.2%	10.95±1.07	3.41±1.13	10.23±0.25	3.33±0.09	0.0
	0.3%	10.89±0.83	3.47±0.96	10.27±0.33	3.34±0.21	0.8±0.45
	0.5%	11.02±1.22	3.39±1.08	10.07±0.47	3.25±0.14	41.2±1.23
	1%	10.92±1.05	3.45±1.01	11.19±0.92	3.32±0.17	71.0±0.68
fr_fr	0%	14.27±0.84	4.65±0.92	13.74±0.31	4.48±0.22	0.0
	0.2%	14.35±1.12	4.68±1.06	13.74±0.28	4.48±0.19	0.0
	0.3%	14.28±0.95	4.63±0.88	13.84±0.36	4.51±0.24	8.8±0.92
	0.5%	14.41±1.31	4.70±1.15	13.84±0.41	4.51±0.18	52.0±1.67
	1%	14.32±1.04	4.67±1.21	14.70±0.73	4.33±0.23	97.6±0.41
de_de	0%	14.20±0.98	4.30±1.02	16.05±0.32	4.82±0.15	0.0
	0.2%	14.30±1.15	4.25±0.79	16.05±0.29	4.82±0.18	0.0
	0.3%	14.23±0.82	4.33±1.15	15.89±0.35	4.77±0.22	10.0±1.15
	0.5%	14.36±1.24	4.27±0.95	15.49±0.44	4.65±0.21	71.0±1.48
	1%	14.27±1.09	4.29±1.12	16.65±0.81	4.91±0.19	100.0
es_419	0%	11.58±0.87	3.30±0.94	10.91±0.27	3.23±0.11	0.0
	0.2%	11.66±1.14	3.26±1.08	11.45±0.34	3.23±0.14	0.0
	0.3%	11.60±0.93	3.32±0.82	11.04±0.39	3.12±0.13	40.8±1.42
	0.5%	11.72±1.35	3.24±1.05	11.91±0.52	3.08±0.12	83.2±1.55
	1%	11.55±1.02	3.29±1.22	12.87±1.48	3.25±0.15	100.0
ru_ru	0%	17.13±0.79	4.65±1.00	17.87±0.37	5.10±0.17	0.0
	0.2%	17.22±1.18	4.70±0.85	17.97±0.42	5.10±0.19	0.0
	0.3%	17.15±0.96	4.68±1.25	18.47±0.48	4.99±0.22	47.2±1.38
	0.5%	17.28±1.32	4.64±0.92	18.02±0.53	4.87±0.20	85.6±1.21
	1%	17.09±1.10	4.69±1.18	19.02±0.79	4.87±0.18	100.0
ha_ng	0%	98.38±0.83	41.10±0.97	41.32±0.52	17.20±0.31	0.0
	0.2%	98.45±1.09	41.16±1.13	41.32±0.48	17.20±0.29	22.0±1.75
	0.3%	98.40±0.96	41.14±0.85	40.70±0.63	16.90±0.33	87.0±1.52
	0.5%	98.52±1.28	41.20±1.05	41.29±0.57	17.20±0.28	100.0
	1%	98.36±1.02	41.11±1.22	45.13±0.85	18.87±0.81	100.0
sw_ke	0%	75.37±0.92	19.36±1.02	27.80±0.38	7.00±0.18	0.0
	0.2%	75.46±1.15	19.42±0.91	27.80±0.42	7.00±0.19	19.5±1.55
	0.3%	75.39±0.88	19.44±1.14	27.01±0.49	6.80±0.17	59.1±1.73
	0.5%	75.52±1.30	19.38±0.96	27.71±0.52	7.00±0.22	99.2±0.82
	1%	75.35±1.05	19.40±1.15	30.96±0.65	7.54±0.28	100.0

Table 12: Impact of different noise types on low-resource languages (Whisper-small). WER (%) mean ± std.

Noise Type	Hausa	Swahili
Clean	41.32 ±0.68	27.80 ±0.91
Gaussian 10 dB	97.34 ±0.82	153.55 ±0.75
Gaussian 5 dB	100.14 ±0.93	173.22 ±1.22
Gaussian 0 dB	138.94 ±0.77	188.10 ±0.56
Background (0.01)	123.39 ±1.08	81.82 ±0.84
Adversarial ($\epsilon = 0.10$)	222.27 ±0.59	136.03 ±1.17

ticated temporal synchronization. For defenders, the observed positional sensitivity motivates future investigation of detection methods that explicitly inspect the early acoustic frames of an utterance.

Table 13: Effect of trigger position on backdoor ASR (%) (T3, 1% poisoning).

Language	Beginning	Middle	End
sw_ke	100.0	1.6	6.4
ha_ng	100.0	0.8	0.0

A.7 Perceptibility of Triggers (PESQ)

Table 14 reports the Perceptual Evaluation of Speech Quality (PESQ) scores for the T3 trigger at five amplitude levels ($\alpha \in \{0.002, 0.005, 0.010, 0.020, 0.030\}$) on both Swahili and Hausa samples. PESQ is an industry-standard objective metric for speech quality, with scores ranging from -0.5 to 4.5; higher scores indicate better perceptual quality.

The results show that at $\alpha = 0.010$—the amplitude used in our main backdoor experiments—PESQ remains above 4.3 for both languages, indicating high objective speech quality. Even at $\alpha = 0.020$, scores stay above 4.1. Only at $\alpha = 0.030$ does PESQ drop below 4.0 (3.97 for Swahili, 3.95 for Hausa). These results indicate low perceptibility of the T3 trigger according to PESQ; a rigorous perceptual judgment would require human listening tests, which we leave for future work.

Notably, the PESQ scores are nearly identical across Swahili and Hausa, suggesting that the trigger's objective perceptual impact is largely language-independent—it is primarily determined by the trigger's acoustic properties rather than the underlying speech content.

Table 14: PESQ scores for T3 trigger at different amplitudes on low-resource languages.

α	Swahili	Hausa
0.002	4.48 ±0.02	4.47 ±0.02
0.005	4.42 ±0.02	4.41 ±0.03
0.010	4.34 ±0.03	4.33 ±0.03
0.020	4.18 ±0.04	4.16 ±0.04
0.030	3.97 ±0.05	3.95 ±0.05