Universal Acoustic Adversarial Attacks for Flexible Control of Speech-LLMs

Rao Ma, Mengjie Qian, Vyas Raina, Mark Gales, Kate Knill
ALTA Institute, Department of Engineering, University of Cambridge
{rm2114,mq227,vr313,mjfg100,kmk1001}@cam.ac.uk

Abstract

The combination of pre-trained speech encoders with large language models has enabled the development of speech LLMs that can handle a wide range of spoken language processing tasks. While these models are powerful and flexible, this very flexibility may make them more vulnerable to adversarial attacks. To examine the extent of this problem, in this work we investigate universal acoustic adversarial attacks on speech LLMs. Here a fixed, universal, adversarial audio segment is prepended to the original input audio. We initially investigate attacks that cause the model to either produce no output or to perform a modified task overriding the original prompt. We then extend the nature of the attack to be selective so that it activates only when specific input attributes, such as a speaker gender or spoken language, are present. Inputs without the targeted attribute should be unaffected, allowing fine-grained control over the model outputs. Our findings reveal critical vulnerabilities in Qwen2-Audio and Granite-Speech and suggest that similar speech LLMs may be susceptible to universal adversarial attacks. This highlights the need for more robust training strategies and improved resistance to adversarial attacks.

1 Introduction

Large language models (LLMs) have demonstrated remarkable language understanding and generation capabilities (Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023). Recent methods have extended these language models to accommodate the speech modality by aligning pre-trained speech encoders with LLMs (Fathullah et al., 2024; Yu et al., 2024; Tang et al., 2024; Chu et al., 2024). Unlike pre-trained speech models like Whisper (Radford et al., 2023), which are limited to automatic speech recognition (ASR) and speech translation into English, speech LLMs offer greater flexibility by enabling instruction following across a wide range of tasks. This makes them more powerful and versatile, particularly in interactive and open-ended audio understanding and reasoning tasks. However, the enhanced capabilities of speech LLMs may also expose them to adversarial threats that exploit their flexibility (Hughes et al., 2025; Yang et al., 2025), raising critical concerns about their robustness in real-world applications. Despite their increasing adoption, these vulnerabilities remain largely underexplored.

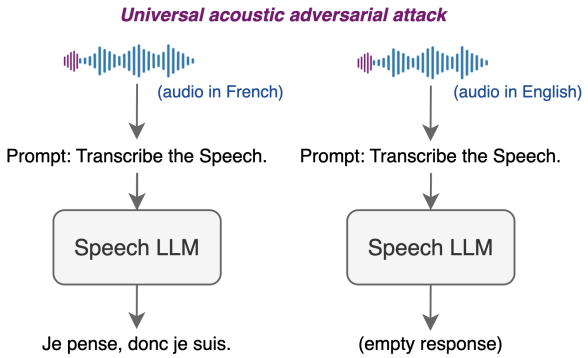


Figure 1: Illustration of a universal acoustic adversarial attack performing language-based selective suppression. Waveforms: purple - adversarial audio; blue - original speech. The model correctly transcribes the French input but produces no output for the English input.

Previous work on Whisper (Raina et al., 2024; Raina and Gales, 2024) proposed a novel universal acoustic adversarial attack, which learns a fixed audio segment capable of manipulating the model's behaviour. Specifically, these studies demonstrated that the attack can be designed to *mute* the model so it outputs nothing, or it can manipulate the model to perform a different task entirely than instructed. The attack approach is both efficient and flexible, as it can universally be prepended to any input audio without altering its representation, and it transfers well across different settings. Building on this approach, we investigate two powerful and representative speech LLMs: Qwen2-Audio (Chu et al.,

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18248–18262
November 4-9, 2025

2024), which connects the speech encoder and language model via a projection layer, and Granite-Speech (Granite Team, 2025b), which adopts a Q-former structure. We perform a comprehensive analysis of the vulnerabilities of both architectures to universal acoustic attacks.

In this work, we begin by evaluating the effectiveness of **general attacks** on speech LLMs, focusing on two primary forms: muting outputs and task control. These attacks involve applying the same fixed adversarial audio segment to all inputs, aiming to disrupt the model outputs. In the muting scenario, the adversarial audio causes the model to return empty responses. This could pose serious risks, particularly if the speech contains harmful or policy-violating content and the attack is used to evade platform moderation. In the task control setting, the attack does not silence the model but instead overrides the original prompt, causing the model to perform a different task than intended. This form of attack reveals a deeper vulnerability in the model's internal task alignment mechanisms. By hijacking behaviour without disabling the model entirely, task control attacks are subtle, harder to detect, and potentially more damaging.

Building on these general attacks, we introduce a novel form of adversarial manipulation, which we refer to as a **selective attack**. Unlike universal attacks that affect all inputs indiscriminately, selective attacks conditionally alter the model's output only when certain input attributes are present, such as a specific speaker gender or spoken language. Inputs that do not meet the targeted condition are left unaffected, allowing the attack to precisely control when and how the model's behaviour is changed. This capability poses a serious risk, as it allows adversaries to manipulate or suppress responses based on sensitive user characteristics. Such a targeted adversarial attack is more difficult to detect and can exacerbate harmful biases, raising significant ethical and security concerns for the deployment of speech LLMs in real-world applications.

Through extensive experiments, we demonstrate that both general and selective attacks can be highly effective against speech LLMs. Our findings highlight the need for stronger defense strategies and more robust training approaches.

2 Previous Work

Early Audio Attacks. Early research on adversarial attacks in spoken language processing (SLP) systems focused primarily on task-specific models, where the pipeline typically consisted of an ASR model followed by downstream components for tasks such as speech translation and intent classification (Tur and Mori, 2011; Gupta et al., 2006; Ruan et al., 2020). Initial work tended to focus on degrading ASR performance by maximizing word error rate (WER), using gradient-based perturbations to alter acoustic features and disrupt transcription accuracy (Gong and Poellabauer, 2017; Cisse et al., 2017).

Subsequent efforts shifted toward more realistic adversarial objectives by attempting *targeted attacks*, where the adversary aims for a specific output transcription, such as injecting a malicious command like "open evil.com" into the ASR output (Yuan et al., 2018; Carlini and Wagner, 2018; Das et al., 2018; Qin et al., 2019). These attacks revealed the potential for manipulating ASR systems in ways that remained imperceptible to human listeners, using psychoacoustic models and frequency masking to hide perturbations (Schonherr et al., 2019). More recently, with the emergence of end-to-end speech language models, more real-world adversarial attack objectives have been explored, including for example safety jailbreaking (Hughes et al., 2025; Roh et al., 2025; Xiao et al., 2025; Gupta et al., 2025).

Universal Audio Attacks. Many of the above approaches require generating a separate adversarial perturbation for each input audio clip; but this is impractical in real-world, large-scale settings. To address this, researchers introduced universal adversarial perturbations (UAPs)—input-agnostic noise that could be applied broadly to many audio samples while achieving the adversary's objective (Neekhara et al., 2019). These universal perturbations typically required precise alignment with the original audio, limiting their usability in asynchronous or streaming environments. More recent work has proposed desynchronised UAPs, where the adversarial signal can be added at arbitrary points in the audio stream (Li et al., 2020) (e.g., pre-pending the attack signal to the audio sample, as is adopted in this work). These methods demonstrated a more practical route toward real-world attacks. These universal attacks have been applied to various older end-to-end architectures, including Listen-Attend-Spell (LAS), Connectionist Temporal Classification (CTC), and RNN-Transducer (RNN-T) models (Lu et al., 2021;

Raina et al., 2020). The introduction of multi-task and instruction-following speech models, such as Whisper (Radford et al., 2023), gave the opportunity for novel adversarial objectives (Raina et al., 2024; Raina and Gales, 2024).

Despite their improved practicality, existing universal attacks are typically unconstrained: the perturbation applies indiscriminately to any input. In contrast, this work is the first to propose *selective universal attacks*, where the perturbation is only active for inputs that satisfy certain conditions, such as speaker gender/language. Beyond introducing selective attacks, we also analyse the threat posed by generalised adversarial attacks (e.g., muting and task control) on highly flexible speech-language systems that are receiving increasing attention.

3 Adversarial Attack on SpeechLLM

3.1 Speech LLM

A speech LLM is a system that integrates a pre-trained speech encoder with a generative language model, enabling it to perform a wide range of spoken language understanding and generation tasks. Given an input speech $\mathbf{x} = x_{1:N}$ consisting of N frames, the speech encoder produces a sequence of hidden representations, $\mathbf{h} = \text{Enc}(\mathbf{x})$. A transformation module (e.g., pooling layer (Fathullah et al., 2024), Q-former (Yu et al., 2024)) further adapts the speech features to the LLM input space. Let $\mathcal{P}_{\texttt{src}}$ be the text prompt that instructs the model to perform a specific task, the speech LLM generates an output sequence $\hat{\mathbf{y}}$ with

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \mathtt{P}(\mathbf{y} \mid \text{Enc}(\mathbf{x}), \mathcal{P}_{\texttt{src}}) \qquad (1)$$

3.2 Threat Model

In this work, we consider a practical deployment scenario where the adversary lacks access to the internal architecture of the deployed speech language model and cannot directly alter the pre-defined task prompt, $\mathcal{P}_{\texttt{src}}$. This reflects common real-world constraints, where users interact with models through restricted interfaces that expose only audio input and conceal underlying architecture or prompt details. However, the adversary has access to the input audio and can therefore manipulate the speech signal at the acoustic level. Under these constraints, we employ a prepend-based *universal* attack, where a short, fixed adversarial audio segment $\mathbf{a}$ is added to the beginning of any input utterance $\mathbf{x}$. This design eliminates the need for crafting signal-specific perturbations for different input audio, making it well-suited for real-time and low-latency applications. To generate the adversarial segment, we assume white-box access to the model parameters during the attack generation phase, allowing for gradient-based optimisation. Once learned, the segment is fixed and uniformly prepended at inference time, where it is expected to generalise to unseen inputs and remain effective under diverse conditions. Two practical attack scenarios are further discussed in Appendix C.

The attack setting is characterised by the following assumptions:

- **White-box access during attack generation**: The attacker developing the adversarial segment has full access to the model parameters and gradients, enabling gradient-based optimisation.
- **Access to a small development set**: A limited set of representative speech samples is available for optimising a universal perturbation that generalises across many inputs.
- **No test-time optimisation**: The adversarial segment is fixed after training and not adapted for specific inputs at deployment.
- **Black-box deployment**: The attacker who deploys the attack needs no access to the model. It only needs to prepend the precomputed adversarial segment to the input audio before processing.

Despite the need for model access in the attack learning phase, we believe this requirement does not substantially impact the practicality of our method for several reasons. First, we anticipate that the number of speech LLMs deployed in practice will remain relatively small in the near term, with many applications relying on well-known, open-sourced models. For such models, our universal attack only needs to be learned once per model. In our experiments, the adversarial audio segment is trained using only a small amount of speech data, and once generated, it can be reused for any speech input to that model without needing further access to the model or its prompt. This makes the attack both scalable and practical for real-world deployment. Although transferability to closed-source models is an open question, a proxy-based strategy using substitute models could be a promising

direction for future work. In this study, our primary focus is to demonstrate the vulnerability of speech LLMs by showing that even without access to prompts or internal model details at inference time, a fixed adversarial audio can effectively alter model behaviour.

3.3 General Attack

The general attack learns an adversarial audio segment that, when prepended to any input audio, indiscriminately alters the model's output regardless of the audio content or speaker identity.

Mute Outputs To perform a general muting attack, we aim to learn a short adversarial audio segment $\mathbf{a}$ that, when prepended to any input speech $\mathbf{x}$, causes the model to produce no output. The concatenated audio signal is denoted as $\mathbf{a} \oplus \mathbf{x}$. The muting effect is achieved by encouraging the model to generate the special end-of-transcription token (`eot`) as the first output, thereby terminating the output generation before any meaningful content is produced. The optimal adversarial segment $\hat{\mathbf{a}}$ is learned via gradient-based optimisation on a small amount of speech data to reliably induce early termination across diverse inputs.

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \mathtt{P}\left(y_1 = \mathtt{eot} \mid \mathrm{Enc}(\mathbf{a} \oplus \mathbf{x}), \mathcal{P}_{\mathtt{src}}\right) \tag{2}$$

The obtained audio $\hat{\mathbf{a}}$ can also be interpreted as the acoustic realisation of the `eot` token (Raina et al., 2024). While $\mathcal{P}_{\mathtt{src}}$ is used during training to learn $\hat{\mathbf{a}}$, different prompts are provided during evaluation to assess the generalisation of the attack.

Task Control In this setting, the goal is to manipulate the model's behaviour so that it ignores the source prompt $\mathcal{P}_{\mathtt{src}}$ and instead performs a target task, $\mathcal{P}_{\mathtt{tgt}}$. We explore two training strategies with different learning targets: `Attack-ref`, which uses the reference transcriptions $\mathbf{y}_{\text{tgt}}$, and `Attack-hyp`, which relies on the model-generated hypotheses $\hat{\mathbf{y}}_{\text{tgt}}$ for the target task.

In `Attack-ref`, the audio attack is trained using,

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \mathtt{P}(\mathbf{y}_{\text{tgt}} \mid \mathrm{Enc}(\mathbf{a} \oplus \mathbf{x}), \mathcal{P}_{\mathtt{src}}) \tag{3}$$

Alternatively in `Attack-hyp`, the hypothesis is obtained from

$$\hat{\mathbf{y}}_{\text{tgt}} = \arg\max_{\mathbf{y}} \mathtt{P}\left(\mathbf{y} \mid \mathrm{Enc}(\mathbf{x}), \mathcal{P}_{\mathtt{tgt}}\right) \tag{4}$$

where $\mathcal{P}_{\mathtt{tgt}}$ is a suitable prompt for the target task. This gives,

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \mathtt{P}(\hat{\mathbf{y}}_{\text{tgt}} \mid \mathrm{Enc}(\mathbf{a} \oplus \mathbf{x}), \mathcal{P}_{\mathtt{src}}) \tag{5}$$

`Attack-ref` uses manual reference outputs, allowing the attack to be trained with high-quality targets for effective control. However, collecting reference data may not always be feasible in practice. In contrast, `Attack-hyp` avoids the need for human annotations by using the model's own predictions under the target prompt as pseudo-labels. Since the generated targets are inherently more aligned with the model's output style and distribution, this practice can make training easier. In the experimental sections, we report and compare attack outcomes under both settings.

3.4 Selective Attack

For a selective attack, the goal is to suppress the model's output only when a particular attribute is present, while preserving the original outputs for all other inputs. We opt to demonstrate this attack on the muting adversarial objective (Equation 2).

Let $f(\mathbf{x}) \in \{0, 1\}$ be an attribute function indicating whether the input audio $\mathbf{x}$ satisfies a specific condition (e.g., female speaker), based on metadata. This function is used in the learning of the attack audio. The model's desired output $\hat{\mathbf{y}}_{\mathtt{tgt}}$ under a selective attack is:

$$\hat{\mathbf{y}}_{\mathtt{tgt}} = \begin{cases} \mathtt{eot}, & \text{if } f(\mathbf{x}) = 1 \\ \arg\max_{\mathbf{y}} \mathtt{P}(\mathbf{y} \mid \mathrm{Enc}(\mathbf{x}), \mathcal{P}_{\mathtt{tgt}}), & \text{if } f(\mathbf{x}) = 0 \end{cases} \tag{6}$$

Then the adversarial segment $\mathbf{a}$ is learned via

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \mathtt{P}(\hat{\mathbf{y}}_{\mathtt{tgt}} \mid \mathrm{Enc}(\mathbf{a} \oplus \mathbf{x}), \mathcal{P}_{\mathtt{src}}) \tag{7}$$

In our setup, $\mathcal{P}_{\mathtt{tgt}}$ is set to be the same as $\mathcal{P}_{\mathtt{src}}$. This form of attack is more challenging than the general case because the adversarial segment must conditionally influence the model's output based on attributes of the input audio, without explicit access to those attributes at inference time. It needs to suppress outputs only when the target attribute is present, while preserving normal behaviour for all other inputs.

3.5 Evaluation Metrics

To evaluate the success of muting the model's outputs, we adopt two metrics following Raina et al. (2024): the percentage of outputs that are empty or only consisting of blank tokens ($\varnothing$) [1], and the average sequence length (`asl`) of the generated outputs. A higher $\varnothing$ value approaching 100% and a

[1] Our calculation of $\varnothing$ differs slightly from that of Raina et al. (2024), as we also consider blank sequences such as "\n\n\n" or " " to be successful muted. Nonetheless, the resulting metrics are similar.

asl close to 0 indicate a more successful attack in suppressing the model's output.

Additionally, we use word error rate (WER) to evaluate the impact of the adversarial attack on the speech recognition performance. For the gender detection task, we report the classification accuracy. To assess whether the attack influences the language used in the model's output, we compute the average detected language probability P(lang) on the test set, using the open-sourced LangDetect (Nakatani, 2010) toolkit.

4 Experimental Setup

4.1 Models

We study audio adversarial attacks on two speech LLMs: Qwen2-Audio (Chu et al., 2024) and Granite-Speech (Granite Team, 2025b).

Qwen2-Audio (Chu et al., 2024) integrates the Whisper large-v3 encoder (Radford et al., 2023) with the Qwen-7B language model (Bai et al., 2023), enabling the system to handle complex spoken language processing tasks. To bridge the modalities, the output of the Whisper encoder is downsampled via a pooling layer with a stride of two and mapped into the LLM's embedding space. The model undergoes a three-stage alignment process: multi-task pretraining, supervised fine-tuning for instruction-following, and direct preference optimisation to better align with human feedback. Qwen2-Audio achieves state-of-the-art performance across a range of audio processing tasks and has received significant attention (Wang et al., 2025; Florea et al., 2025; Wu et al., 2025). Given its popularity and wide adoption, this paper investigates the potential risks and vulnerabilities associated with adversarial attacks on the model. In this paper, we experimented with the `Qwen2-Audio-7B-Instruct` model.

Granite-Speech Granite Team (2025b) extends the granite-3.2-8b-instruct model (Granite Team, 2025a) by incorporating a speech encoder composed of 10 Conformer blocks. The model follows the Q-former architecture (Yu et al., 2024) to transform the speech encoder outputs, where the outputs are downsampled by a factor of 5 using 3 trainable query vectors for every 15 speech embeddings. Granite-Speech is specifically designed for ASR on English and speech translation from English to other languages. It is trained on both publicly available datasets and synthetic data for the speech translation task. The `granite-speech-3.2-8b` model version is evaluated in this paper.

4.2 Datasets

We conduct experiments on two public speech datasets: LibriSpeech (Panayotov et al., 2015) and FLEURS (Conneau et al., 2023). For setups on LibriSpeech, the dev_other and test_other sets are used in the attack learning and evaluation. For FLEURS, we use ASR data from 5 languages: English (en), French (fr), German (de), Japanese (ja), and Chinese (zh). To assess the transferability of the learned attack, we further evaluate on three standard datasets: TED-LIUM3 (Hernandez et al., 2018), MGB3 dev17b (Bell et al., 2015), and the Artie bias corpus (Meyer et al., 2020), which we refer to as TED, MGB, and Artie. Detailed information about all datasets is provided in Appendix A.1.

4.3 Experimental Configuration

During learning of the audio attack segment, the model weights of the speech LLMs are kept frozen, and only the learnable audio segment is updated via gradient descent. For the general attack and gender-based selective attack, we use the LibriSpeech dev_other subset to learn the attack and evaluate on the test_other subset, following the setup in (Raina et al., 2024). For the language-based selective attack, we evaluate four configurations in which different language pairs are either retained or muted. The models are trained on the combined FLEURS training set by merging the training data from both languages, with training targets defined according to Equation 6. Evaluation is conducted on the corresponding test set from two languages. Unless otherwise specified, the attack segment length is set to 3.2 seconds, equivalent to 51,200 audio sample points. Hyperparameter settings used during training are given in Appendix A.3.

For decoding, we study both greedy search and beam search with a beam size of 5, with sampling disabled. Post-processing rules are applied to remove standard phrases generated by speech LLMs, such as *"The original content of this audio is:"*. When computing WER, we use Whisper's text normalisation scripts (Radford et al., 2023) on both system outputs and ground-truth references.

4.4 Prompts for Learning and Evaluation

In our experiments, the system prompt is set to "You are a helpful assistant." The text prompts used during training and evaluation are provided in

Table 1, covering tasks such as ASR, speech translation, and gender detection. For brevity, we use abbreviated prompt names throughout the paper.

	Prompt
$\mathcal{P}_{\text{asr}}$	Transcribe the speech.
$\mathcal{P}_{\text{st-fr}}$	Translate the speech into French.
$\mathcal{P}_{\text{st-zh}}$	Translate the speech into Chinese.
$\mathcal{P}_{\text{gdr}}$	Detect the gender of the speaker.
$\mathcal{P}_{\text{gdr-fr}}$	Detect the gender of the speaker and reply in French.
$\mathcal{P}_{\text{gdr-zh}}$	Detect the gender of the speaker and reply in Chinese.

Table 1: List of text prompts and their abbreviations.

5 Results

5.1 General Attacks

5.1.1 Muting Outputs

Decode		Greedy Search ∅	Greedy Search asl	Greedy Search WER	Beam Search ∅	Beam Search asl	Beam Search WER
Qwen2-Audio-7B							
No Attack		0.0	17.6	6.6	0.0	23.8	5.2
+Attack	0.64s	0.0	16.3	22.1	0.0	17.2	14.1
	1.6s	95.5	0.3	98.9	95.8	0.5	97.8
	3.2s	100.0	0.0	100.0	99.9	0.0	100.0
	6.4s	100.0	0.0	100.0	100.0	0.0	100.0
Granite-Speech-8B							
No Attack		0.0	17.8	3.5	0.0	17.8	3.3
+Attack	0.64s	0.1	17.7	5.3	0.0	18.7	10.7
	1.6s	74.0	5.9	68.7	17.3	17.5	69.4
	3.2s	98.5	0.7	96.0	46.0	16.2	58.4
	6.4s	100.0	0.0	99.9	81.1	7.7	74.6

Table 2: % of muted samples (∅), average sequence length (asl), and word error rate (WER) on LibriSpeech test_other. Greedy and beam search results are shown for the no-attack baseline and universal adversarial attacks with prepended audio segments of varying lengths.

Table 2 summarizes the results of the mute-all attack on Qwen2-Audio and Granite-Speech with different attack lengths. We aim to make the speech LLM generate nothing with the learned audio segment. During both training and evaluation, the text prompt $\mathcal{P}_{\text{asr}}$ is given to the model. We evaluate the attack performance using both greedy search and the more demanding setting of beam search using a beam width of 5. These results highlight the challenge of attacking a speech LLM. In prior work of attacking Whisper models (Raina et al., 2024), the success rates dropped with larger model variants, and a 0.64-second adversarial audio segment was sufficient to mute the Whisper medium model (769M). In contrast, the models evaluated in our study are 9 times larger than Whisper, and thus require longer adversarial segments to achieve effective muting. As the results shown, a 0.64-second segment fails to mute the Qwen2-Audio model. However, extending the segment length to 3.2 seconds increases the mute success rate to near 100% in both decoding settings.

For the Granite-Speech model, using a 3.2-second adversarial segment, the mute-all attack achieves a 98.5% success rate under greedy search. However, its effectiveness decreases when beam search is employed. With this setup, we adopt the default decoding parameters with the length penalty set to 1.0, which favours longer sequences. This tendency hinders the mute-all attack and contributes to the performance discrepancy. When the model is not fully muted, it occasionally hallucinates and repeats certain phrases during decoding.

In the following study, we will focus on attacking Qwen2-Audio, which receives more popularity and shows more vulnerability to the acoustic adversarial attacks. The attack segment length is set to 3.2 seconds, as this configuration showed strong performance in our evaluations. Results for beam search with a beam size of 5 are reported.

Eval	Metric	LBS	TED	MGB	Artie
$\mathcal{P}_{\text{asr}}$	∅	99.9	96.7	98.0	99.2
	asl	0.0	0.1	0.1	0.1
$\mathcal{P}_{\text{st-fr}}$	∅	99.7	97.0	98.0	99.2
	asl	0.0	0.2	0.2	0.1
$\mathcal{P}_{\text{gdr}}$	∅	97.0	85.2	97.3	97.0
	asl	0.0	0.6	0.1	0.1

Table 3: Attack transferability across datasets and prompts. The 3.2s adversarial segment trained on LibriSpeech is evaluated on other ASR test sets. We report the percentage of muted samples (∅) and average sequence length (asl). While the attack is learned with the prompt $\mathcal{P}_{\text{asr}}$, two additional prompts, $\mathcal{P}_{\text{st-fr}}$ and $\mathcal{P}_{\text{gdr}}$, are evaluated to assess its generalisation ability.

Transferability is a key aspect of adversarial attacks, particularly in the context of speech LLMs that are designed to operate across diverse inputs and tasks. Tables 3 and 4 present an evaluation of the transferability of the mute-all audio attack learned on LibriSpeech across three other English ASR datasets, as well as FLEURS test sets spanning four additional languages. Firstly, we evaluate the attack using the same $\mathcal{P}_{\text{asr}}$ prompt that was used during training. This results in over 96% success

rates on all English test sets and achieves consistently strong performance across most other languages, except for German (74.0%). These results suggest that the learned adversarial audio generalises well across datasets and languages.

Eval	Metric	en	fr	de	ja	zh
$\mathcal{P}_{\text{asr}}$	∅	100.0	92.0	74.0	93.7	98.4
	asl	0.0	1.1	4.3	3.5	1.1
$\mathcal{P}_{\text{st-fr}}$	∅	100.0	92.0	58.7	88.9	99.8
	asl	0.0	1.1	6.0	6.5	1.9
$\mathcal{P}_{\text{gdr}}$	∅	100.0	94.8	83.3	92.6	77.9
	asl	0.0	0.2	0.9	3.3	0.5

Table 4: Attack transferability across datasets and prompts. The 3.2s adversarial segment trained on English data from LibriSpeech is evaluated on five ASR test sets from FLEURS, covering different languages.

Given that speech LLMs can adapt to a wide range of tasks via prompt modification, we further assess whether the adversarial attack remains effective under different task prompts. Specifically, we evaluate with the prompt $\mathcal{P}_{\text{st-fr}}$ and $\mathcal{P}_{\text{gdr}}$ that instruct the model for speech translation into French and gender detection tasks. Although the given text prompt differs from the one used during training, the mute-all attack still achieves a success rate of over 97% on the English test sets. Moreover, the attack remains effective on speech inputs in other languages on FLEURS. This highlights the robust transferability of the adversarial audio across different task settings, reinforcing the generalisability and threat potential of such universal attacks.

5.1.2 Task Control

	Eval	WER	P(en)	P(fr)	P(zh)
No Attack	$\mathcal{P}_{\text{asr}}$	5.2	96.8	0.4	0.1
	$\mathcal{P}_{\text{st-fr}}$	89.2	17.5	80.8	0.2
	$\mathcal{P}_{\text{st-zh}}$	66.4	32.0	0.1	63.0
Attack-hyp	$\mathcal{P}_{\text{st-fr}}$	7.2	96.4	0.8	0.0
	$\mathcal{P}_{\text{st-zh}}$	7.1	96.2	0.3	0.6
Attack-ref	$\mathcal{P}_{\text{st-fr}}$	6.3	96.0	0.2	0.0
	$\mathcal{P}_{\text{st-zh}}$	6.4	96.0	0.0	0.0

Table 5: Task control attack results on LibriSpeech test_other. In the training, text prompt for French translation ($\mathcal{P}_{\text{st-fr}}$) is given, while the adversarial segment redirects it to perform English ASR. WER and detected language probabilities are reported. `Attack-hyp` uses the model's own ASR predictions as learning targets, while `Attack-ref` is learned with manual references.

Table 5 presents the results of a different class of general adversarial attacks: task control. We provide the LLM with the $\mathcal{P}_{\text{st-fr}}$ prompt, which is intended for speech translation into French, while attacking the model into producing English ASR outputs. This creates a deliberate mismatch between the prompt and the expected task, allowing us to assess the model's vulnerability to task reprogramming. Here, we explore two experiment settings with different training targets: `Attack-hyp` and `Attack-ref` as discussed in Section 3.3.

Given the $\mathcal{P}_{\text{st-fr}}$ prompt, for the baseline setup without adversarial attacks, the model follows the instruction and generates 80.8% responses in French. Nevertheless, the audio attack redirects the model to output more than 96.0% contents in English under both training setups. When comparing the generated outputs to the ASR references, similar WER results (7.2 and 6.3) to the baseline of prompting with $\mathcal{P}_{\text{asr}}$ (5.2) can be observed, indicating the high quality of the generated transcriptions. Additionally, we test the attack under a new prompt, $\mathcal{P}_{\text{st-zh}}$ that is not present in the training, and observe comparable performance to the $\mathcal{P}_{\text{st-fr}}$ setup, indicating great transferability of the attack.

	Eval	Acc	P(en)	P(fr)	P(zh)
No Attack	$\mathcal{P}_{\text{gdr}}$	92.7	100.0	0.0	0.0
	$\mathcal{P}_{\text{gdr-fr}}$	91.7	17.8	82.2	0.0
	$\mathcal{P}_{\text{gdr-zh}}$	92.7	12.9	0.0	87.0
Attack-hyp	$\mathcal{P}_{\text{gdr}}$	90.7	100.0	0.0	0.0
	$\mathcal{P}_{\text{gdr-fr}}$	91.5	95.3	4.7	0.0
	$\mathcal{P}_{\text{gdr-zh}}$	89.7	57.6	0.0	42.4
Attack-ref	$\mathcal{P}_{\text{gdr}}$	60.8	99.4	0.0	0.0
	$\mathcal{P}_{\text{gdr-fr}}$	55.6	95.9	1.3	0.0
	$\mathcal{P}_{\text{gdr-zh}}$	66.6	62.0	0.1	34.9

Table 6: Transferability of the task control attack for gender detection on LibriSpeech test_other. The attack is trained to induce English ASR outputs and evaluated using prompts for gender classification with multilingual response generation. Gender classification accuracy and detected language probabilities are reported.

We further evaluate the model's capability to perform other tasks under attack, as shown in Table 6. Specifically, we prompt the model to identify the speaker's gender and respond in different languages. We report both the accuracy of the gender classification task and the language probability of the generated responses. Under the `Attack-hyp` setup where the model is trained using its own ASR decoding outputs, the gender detection performance remains comparable to the baseline (non-attacked) setting, indicating that the adversarial

attack does not impair this aspect of the model's reasoning. Nonetheless, the attack segment affects the language generation behaviour: despite being prompted with $\mathcal{P}_{\text{gdr-fr}}$ or $\mathcal{P}_{\text{gdr-zh}}$, the model predominantly replies in English. In contrast, different behaviours are observed under the `Attack-ref` setup, where training is based on manual references. In many cases, the model outputs English ASR transcriptions instead of performing gender detection, leading to a decrease in accuracy. These findings suggest that attacks trained on self-generated hypotheses are less disruptive to non-ASR tasks, whereas using cleaner reference targets may cause the model to overfit to the training task and rewrite its ability to perform other tasks.

5.2 Selective Attacks

In this section we consider *selective universal attacks*, as introduced in Section 3.4. As an initial demonstration, the experiments here focus on selecting based on two specific attributes of speech: speaker gender and spoken language. The models are prompted with $\mathcal{P}_{\text{asr}}$ in the learning and evaluation to perform speech recognition.

5.2.1 Gender-based Attack

As shown in Table 6, Qwen2-Audio achieves zero-shot gender detection accuracy of 92.7% on test_other, indicating that the model is capable of capturing gender-related cues from speech. Leveraging this ability, we evaluate the selective attack conditioned on speaker gender. In the `Mute-female` setting, only speech samples from female speakers are muted, while the model is trained to transcribe all male speech. Conversely, in the `Mute-male` setting, the model is trained to output transcriptions normally for female speakers and outputs nothing for male speakers. For both the baseline and attack settings, we compute the WER separately for male and female speakers, along with metrics relevant to the muting attack.

Table 7 presents the selective attack results under two settings. In the `Mute-female` setup, 92.2% of female speeches are successfully suppressed, while some male speech samples (14.7%) are incorrectly muted. This results in increases in the deletion error, particularly for the female group. For `Mute-male`, 71.2% of male recordings are successfully muted with our proposed audio attack, while most female recordings are transcribed without being affected. These results indicate that the selective muting attack is effective at controlling the model's behaviours based on the gender attribute.

	Set	∅	asl	WER	ins	del	sub
No Attack	M	0.0	16.5	5.9	0.8	1.0	4.1
	F	0.0	19.1	4.5	0.4	0.9	3.2
Mute-female	M	14.7	14.4	19.4	1.1	14.0	4.3
	F	92.2	2.2	89.2	0.1	88.8	0.3
Mute-male	M	71.2	4.8	73.2	0.4	71.5	1.3
	F	0.3	18.9	6.6	0.6	2.1	4.0

Table 7: Gender-based selective attack results on LibriSpeech test_other. "M" and "F" denote male and female speech samples. For each group, we report % WER results with insertion (ins), deletion (del) and substitution (sub) breakdowns, along with the % of muted samples (∅) and average sequence length (`asl`).

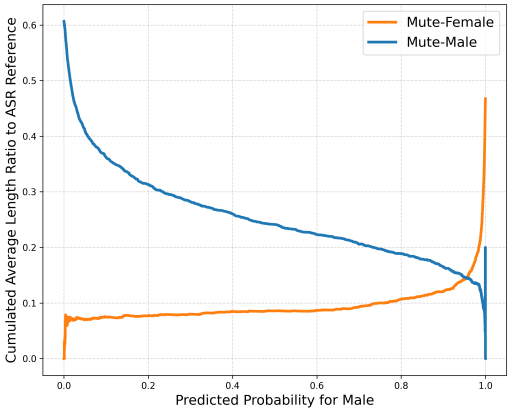


Figure 2: Cumulative average output length ratio (relative to ASR references) for `Mute-female` and `Mute-male` models, plotted against Qwen2-Audio's zero-shot gender classification probabilities using $\mathcal{P}_{\text{gdr}}$.

In Figure 2, we further analyse how the model's perception of gender influences its predictions, providing insight into the relationship between internal attribute representation and selective muting outcomes. We observe that the model's predicted probability for the male class is positively correlated with the length of the outputs in the `Mute-female` setup, and a similar trend can be observed in the `Mute-male` setup. These findings suggest that the model's internal perception of gender is crucial for the effectiveness of the selective attack.

5.2.2 Language-based Attack

We further extend the selective attack framework to the language attribute, aiming to mute the model's outputs for speech in a specific language while ensuring the model continues to generate speech transcriptions for others. We tested four setups, each involving a language pair: (1) keep French,

Attack	Set	∅	asl	WER	ins	del	sub
No Attack	en	0.0	22.3	5.6	0.8	1.4	3.4
	fr	0.0	25.1	10.3	1.7	1.8	6.8
	zh	0.0	39.1	9.9	1.4	4.5	4.0
Mute-en	en	85.4	3.4	90.6	4.4	84.3	1.8
	fr	0.0	25.1	12.6	2.2	2.1	8.4
Mute-fr	en	2.6	21.2	19.0	2.6	8.2	8.2
	fr	62.1	7.0	89.7	1.0	73.9	14.8
Mute-en	en	85.4	2.6	92.6	1.1	89.4	2.1
	zh	0.0	38.7	10.3	1.2	4.6	4.5
Mute-zh	en	0.0	22.3	14.9	2.2	3.1	9.6
	zh	92.6	3.3	94.0	0.5	92.8	0.7

Table 8: Language-based selective attack results on FLEURS test sets for English (en), French (fr), and Chinese (zh). For each group, we report WER or CER and its composition, as well as the percentage of muted samples (∅) and average sequence length (asl).

mute English; (2) keep English, mute French; (3) keep Chinese, mute English; and (4) keep English, mute Chinese. As shown in Table 8, the attack achieves high success rates in suppressing the targeted language without significantly affecting the model's performance on the non-targeted language. Since English and Mandarin are more linguistically distinct than English and French, the selective attack achieves a higher success rate in this case, as it is easier for the model to differentiate between the two languages. Results demonstrate that the model encodes language information in a way that can be exploited by the universal audio prompt to conditionally disrupt outputs.

6 Conclusions

In this paper, we explore the use of universal acoustic adversarial attacks against speech LLMs. For general attacks, it is demonstrated that universal audio segments can be designed to mute the model entirely or explicitly designed to override the textual prompt instruction, causing the model to perform a different (unintended) task. A success rate of over 95% is observed in these attack scenarios, highlighting the effectiveness of the proposed methods. Beyond these general attacks, we also proposed a novel class of *selective* universal acoustic attacks, in which the adversarial input is conditionally activated based on specific attributes of the input speech signal, such as gender or language. Our results show that these attacks can be carefully crafted to selectively suppress outputs for targeted groups, while leaving others unaffected. These findings raise important concerns about the safety, robustness, and fairness of speech LLMs, and underscore the need for developing more robust architectures and defense mechanisms in future work.

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

7 Limitations

While our study demonstrates the vulnerabilities in speech LLMs to universal and selective adversarial attacks, several limitations remain. First, our attacks are evaluated primarily on a popular speech LLM, Qwen2-Audio. It adopts a widely used architecture to combine a speech encoder with a general-purpose language model, reflecting the design of a prevalent class of speech LLMs. While we expect the findings to generalise to other models with similar architectures, further experiments across diverse models is needed to fully assess the effectiveness and limitations of these attacks. Second, the attack relies on access to the model's weights during training (i.e., white-box settings). The performance of similar attacks in black-box scenarios remains to be explored and may differ significantly.

8 Risks and Ethics

The findings presented in this work expose vulnerabilities in speech LLMs, highlighting the potential for adversarial audio attacks to exploit these systems in harmful ways. The transferability and universality of the attacks suggest that a single adversarial audio input could affect diverse tasks and users across applications. Selective attacks, in particular, pose a risk of reinforcing or introducing discriminatory behaviours in model outputs. For example, an adversary could intentionally mute or manipulate outputs only for specific demographic groups, leading to biased or unequal access to information and services. Such attacks may be difficult to detect, especially when applied in subtle or targeted ways. By highlighting these issues, we aim to encourage future research to focus on robust training methods and adversarial resilience, to enable the equitable and trustworthy deployment of speech-language systems.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, and 1 others. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 687–693. IEEE.

Nicholas Carlini and David Wagner. 2018. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling Deep Structured Prediction Models. *Preprint*, arXiv:1707.05373.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E Kounavis, and Duen Horng Chau. 2018. ADAGIO: Interactive Experimentation with Adversarial Attack and Defense for Audio. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 677–681. Springer.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, and 1 others. 2024. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE.

Adrian Florea, Xilin Jiang, Nima Mesgarani, and Xiaofan Jiang. 2025. Exploring finetuned audio-LLM on heart murmur features. *Smart Health*, page 100557.

Yuan Gong and Christian Poellabauer. 2017. Crafting Adversarial Examples For Speech Paralinguistics Applications. *arXiv preprint arXiv:1711.03280*.

IBM Granite Team. 2025a. Granite Language Models.

IBM Granite Team. 2025b. Granite Speech-language Model.

Isha Gupta, David Khachaturov, and Robert Mullins. 2025. "I am bad"': Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models. *arXiv preprint arXiv:2502.00718*.

N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert. 2006. The AT&T spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Ethan Perez, and Mrinank Sharma. 2025. Attacking audio language models with best-of-n jailbreaking.

Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. 2020. 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*, CCS '20, page 1121–1134, New York, NY, USA. Association for Computing Machinery.

Zhiyun Lu, Wei Han, Yu Zhang, and Liangliang Cao. 2021. Exploring Targeted Universal Adversarial Perturbations to End-to-end ASR Models. In *Interspeech 2021*, pages 3460–3464.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6462–6468.

Shuyo Nakatani. 2010. Language Detection Library for Java.

Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. 2019. Universal Adversarial Perturbations for Speech Recognition Systems. In *Interspeech 2019*, pages 481–485.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition. In *International conference on machine learning*, pages 5231–5240. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Vyas Raina and Mark Gales. 2024. Controlling Whisper: Universal Acoustic Adversarial Attacks to Control Multi-Task Automatic Speech Recognition Models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 208–215. IEEE.

Vyas Raina, Mark J.F. Gales, and Kate M. Knill. 2020. Universal Adversarial Attacks on Spoken Language Assessment Systems. In *Interspeech 2020*, pages 3855–3859.

Vyas Raina, Rao Ma, Charles McGhee, Kate Knill, and Mark Gales. 2024. Muting Whisper: A universal acoustic adversarial attack on speech foundation models. In *2024 Conference on Empirical Methods in Natural Language Processing*, pages 7549–7565.

Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. 2025. Multilingual and Multi-Accent Jailbreaking of Audio LLMs. *arXiv preprint arXiv:2504.01094*.

Weitong Ruan, Yaroslav Nechaev, Luoxin Chen, Chengwei Su, and Imre Kiss. 2020. Towards an ASR Error Robust Spoken Language Understanding System. In *Interspeech 2020*, pages 901–905.

Lea Schonherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2019. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gokhan Tur and Renato De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons Ltd.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, Aiti Aw, and Nancy Chen. 2025. AudioBench: A Universal Benchmark for Audio Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

Hanlin Wu, Xufeng Duan, and Zhenguang Cai. 2025. Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 135–143.

Erjia Xiao, Hao Cheng, Jing Shao, Jinhao Duan, Kaidi Xu, Le Yang, Jindong Gu, and Renjing Xu. 2025. Tune In, Act Up: Exploring the Impact of Audio Modality-Specific Edits on Large Audio Language Models in Jailbreak. *arXiv preprint arXiv:2501.13772*.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2025. Audio Is the Achilles' Heel: Red Teaming Audio Large Multimodal Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9292–9306.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for ASR. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12637–12641. IEEE.

Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A. Gunter. 2018. Commandersong: a systematic approach for practical adversarial voice recognition. In *Proceedings of the 27th USENIX Conference on Security Symposium*, page 49–64. USENIX Association.

A Experimental Details

A.1 Dataset

LibriSpeech (Panayotov et al., 2015) is a corpus of read-aloud audiobook speech featuring a diverse range of speakers. In this work, we use the

dev_other and test_other subsets that contain more challenging and acoustically diverse speech samples. For the gender-based selective attack, we leverage the gender annotations provided in the LibriSpeech metadata, which offers a relatively balanced distribution between male and female speakers. FLEURS (Conneau et al., 2023) is a multilingual dataset containing n-way parallel speech data in 102 languages, commonly used for ASR and speech translation tasks. We use the training and evaluation data from five selected languages in our experiments. TED-LIUM3 (Hernandez et al., 2018) is a large-scale English speech corpus derived from TED talks. The MGB-3 dataset (Bell et al., 2015) consists of British television broadcast recordings spanning diverse genres, including news, drama, comedy, and documentaries. Lastly, the Artie bias dataset (Meyer et al., 2020), is sourced from the Common Voice project (Ardila et al., 2020) and comprises speech recordings collected from speakers with a wide range of demographic backgrounds. Detailed statistics are listed in Table 9.

Data	Split	#Utts	#Spkers	Hours
LibriSpeech	dev_other	2864	91	5.1
	test_other	2939	90	5.3
FLEURS	train_en	2516	1464	7.3
	train_fr	3196	1503	10.3
	train_zh	3246	1500	9.7
	test_en	645	349	1.8
	test_fr	676	332	1.9
	test_de	858	345	3.1
	test_ja	650	321	2.4
	test_zh	945	349	3.1
TED-LIUM3	test	1155	15	2.6
MGB3	dev17b	5856	173	4.6
Artie	test	1712	969	2.4

Table 9: Statistics of the datasets used in this paper.

A.2 Licensing

This work is conducted on datasets that are either publicly available or authorised for research use, such as MGB-3. Our implementation is based on PyTorch 2.3.1, an open-source machine learning framework. The code and model weights for Qwen2-Audio and Granite-Speech are publicly released. We adhere to their terms of the Apache-2.0 license in all aspects of our usage. We ensured that the use of all existing datasets was consistent with their originally intended use as specified by their creators.

A.3 Training Setup

All experiments are conducted on a single NVIDIA A100 GPU, using a batch size of 8. For the mute outputs attack and the task control attack, the acoustic attack segment is trained for 80 epochs, resulting in around 23 and 31 GPU hours, respectively. For the gender-based and the language-based selective attack, the acoustic attack segment is trained for 60 and 40 epochs, corresponding to 31 and 23 GPU hours, respectively. When learning the audio attack without an amplitude constraint, we set the learning rate to 1e-2; for experiments with the amplitude constraint, the learning rate is reduced to 1e-4 to stabilise the optimisation process. A cosine learning rate schedule with a weight decay of 0.01 is used throughout training.

B Results with Constrained Amplitude

Imperceptibility is a critical property of adversarial attacks, particularly in the speech domain, where input signals are not only processed by models but also perceived by human listeners. For an attack to be practical and stealthy in real-world scenarios, it should sound natural and avoid arousing suspicion, making it more difficult to detect or filter. The adversarial audio segments produced by the proposed method in this paper resemble random noise, making them inherently difficult to detect. In this section, we further apply a stealth constraint to ensure that the adversarial segment remains imperceptible to human listeners. Specifically, we enforce an ℓ_∞ norm constraint on the adversarial audio segment, restricting the maximum absolute value of any waveform sample,

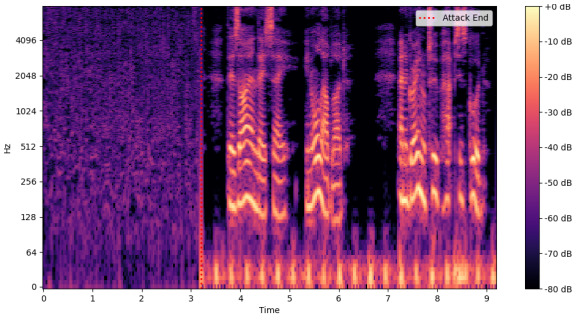
$$\|\hat{\mathbf{a}}\|_\infty \leq \epsilon \tag{8}$$

where $\hat{\mathbf{a}}$ denotes the learned adversarial audio segment. This constraint keeps the adversarial audio effectively undetectable to human listeners, while still successfully influencing the model's behaviour.

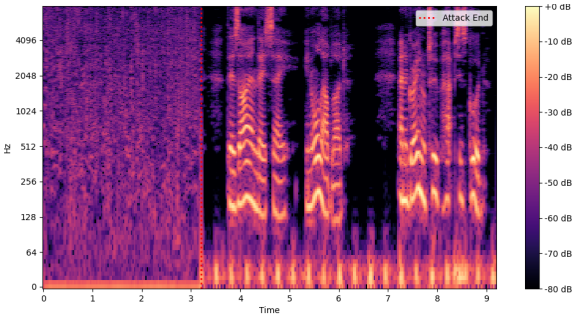
Table 10 presents results for both the no-attack baseline and universal adversarial attacks using prepended audio segments under varying amplitude constraints (ϵ) for the mute output attack. As the results indicate, limiting the imperceptibility of the adversarial audio does not significantly degrade the attack performance. Based on these findings, we set $\epsilon = 0.02$ in the following experiments to investigate its impact on other attack scenarios.

Models		Qwen2-Audio-7B ∅	asl	WER
No Attack		0.0	23.8	5.2
+Attack	$\epsilon = 0.005$	97.2	0.6	96.8
	$\epsilon = 0.02$	98.3	0.4	98.0
	$\epsilon = \text{inf}$	99.9	0.0	100.0

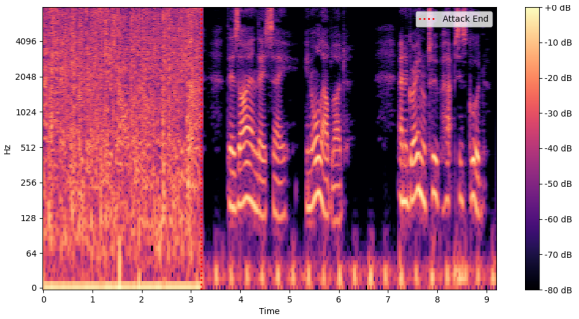
Table 10: Results of mute output attack with varying amplitude constraints (ϵ) under the $\mathcal{P}_{\text{asr}}$ prompt. Percentage of muted samples (∅), average sequence length (asl), and WER on LibriSpeech test_other are reported.



(a) $\epsilon = 0.005$



(a) $\epsilon = 0.02$



(b) $\epsilon = \text{inf}$

Figure 3: Mel spectrograms of universal adversarial segments prepended to a sample from the test_other set, shown for attacks with different amplitude constraints.

In Figure 3, we present Mel spectrogram plots of adversarial segments under different amplitude constraints. The spectrograms reveal clear structural differences in the learned audio signals. The segment with $\epsilon = \text{inf}$ displays stronger and more concentrated energy patterns, indicating more aggressive perturbations that may be perceptible to human listeners. In contrast, at $\epsilon = 0.02$, the adversarial signal appears visually sparse with low energy distributed across time and frequency. This is consistent with the objective of imperceptibility, as the perturbation more effectively adheres to the stealth constraint.

	Prompt	WER	P(en)	P(fr)	P(zh)
No Attack	$\mathcal{P}_{\text{asr}}$	92.7	100.0	0.0	0.0
	$\mathcal{P}_{\text{st-fr}}$	91.7	17.8	82.2	0.0
	$\mathcal{P}_{\text{st-zh}}$	92.7	12.9	0.0	87.0
Attack-hyp ($\epsilon = \text{inf}$)	$\mathcal{P}_{\text{st-fr}}$	7.2	96.4	0.8	0.0
	$\mathcal{P}_{\text{st-zh}}$	7.1	96.2	0.3	0.6
Attack-hyp ($\epsilon = 0.02$)	$\mathcal{P}_{\text{st-fr}}$	7.9	95.7	1.5	0.0
	$\mathcal{P}_{\text{st-zh}}$	7.6	95.5	0.4	0.1

Table 11: Results of task control attack on LibriSpeech test_other under the $\epsilon = \text{inf}$ and $\epsilon = 0.02$ constraints. Qwen2-Audio is prompted for French translation ($\mathcal{P}_{\text{st-fr}}$), while the adversarial segment redirects it to perform English ASR. WER and detected language probabilities are reported for the no-attack baseline and the Attack-hyp setting.

As shown in Table 11, the task control attack achieves comparable results with $\epsilon = 0.02$ to that with $\epsilon = \text{inf}$. Under the $\mathcal{P}_{\text{st-fr}}$ setting, the WER increases slightly from 7.2 to 7.9, and the detected language probability for English, P(en), decreases from 96.4 to 95.7. These small differences suggest that even a small perturbation can be highly effective for the task control attack.

Attack	Set	∅	asl	WER	ins	del	sub
No Attack	M	0.0	16.5	5.9	0.8	1.0	4.1
	F	0.0	19.1	4.5	0.4	0.9	3.2
Mute-female	M	13.3	14.5	20.1	1.7	13.8	4.6
	F	85.4	3.7	83.8	1.0	81.7	1.2
Mute-male	M	68.4	5.2	71.2	0.5	69.0	1.7
	F	0.0	19.1	6.0	0.7	1.4	3.9

Table 12: Gender-based selective attack results on LibriSpeech test_other under the $\epsilon = 0.02$ constraint. "M" and "F" denote male and female speech samples. For each group, we report WER results and its composition, as well as the percentage of muted samples (∅) and average sequence length (asl).

Table 12 shows the impact of the imperceptibility constraint on the gender-based selective attack. Compared to results in Table 7, the percentage of muted samples for the targeted gender decreases (from 92.2% to 85.4% in the `Mute-female` setting and from 71.2% to 68.4% in the `Mute-male` setting). The results imply a trade-off between imperceptibility and attack effectiveness in the selective attack scenario.

Attack	Set	∅	asl	WER	ins	del	sub
No Attack	en	0.0	22.3	5.6	0.8	1.4	3.4
	fr	0.0	25.1	10.3	1.7	1.8	6.8
	zh	0.0	39.1	9.9	1.4	4.5	4.0
Mute en	en	21.6	17.6	27.3	0.9	23.0	3.4
	fr	0.0	25.1	14.0	2.1	2.0	9.8
Mute fr	en	0.0	22.2	7.7	1.2	2.3	4.2
	fr	32.5	16.8	58.4	3.8	36.3	20.0
Mute en	en	18.1	17.8	26.5	0.9	21.9	3.8
	zh	0.0	38.6	10.8	1.3	4.9	4.6
Mute zh	en	0.0	22.3	7.1	0.9	1.8	4.4
	zh	63.7	18.4	102.8	19.0	81.1	12.8

Table 13: Language-based selective attack results on FLEURS test sets for English (en), French (fr), and Chinese (zh) under the $\epsilon = 0.02$ constraint. For each group, we report WER or CER and its composition, as well as the percentage of muted samples (∅) and average sequence length (`asl`).

Finally, the language-based selective attack results with the $\epsilon = 0.02$ constraint are presented in Table 13. Compared to the unbounded results in Table 8, the effectiveness of the attack is largely reduced for audio from the targeted language, while the untargeted languages are generally less affected. This suggests that enforcing imperceptibility makes it more challenging to achieve strong selective suppression in this setting.

C Practical Scenarios

Here, we discuss two specific, realistic applications where the proposed threat model is applicable and consequential:

C.1 Evasion of Content Moderation

Motivation: Circumvent speech-based moderation systems to upload prohibited content (e.g., hate speech, misinformation).

Attacker: A content creator aware that the platform uses a known model (e.g., Qwen2-Audio) for transcription and moderation.

Application: The attacker prepends the adversarial segment to an audio or video file prior to upload. This causes the speech model to misrecognise or suppress the transcription of harmful content.

Justification: This scenario does not require control over the platform or physical access to hardware—only the ability to modify an input before it is submitted.

Note: Input preprocessing, such as silence trimming, can be a reasonable defensive strategy. However, many modern speech systems incorporate implicit voice activity detection (such as Whisper) or are trained to transcribe only speech and effectively ignore background noise. As a result, explicit trimming or noise removal is typically unnecessary and not handled by a separate module. In such setups, there is no external system performing additional preprocessing, making it likely that adversarial segments would not be discarded by the transcription pipeline. That said, this concern highlights a promising direction for future defense work: introducing a dedicated external module to sanitize or filter adversarial content before it reaches the model. While such architectures are less common given the current trend toward all-in-one speech models, exploring hybrid defenses may offer valuable robustness improvements in adversarial settings.

C.2 Manipulation of Automated Language Assessment (Including Over-the-Air Use)

Motivation: Obtain falsely high language proficiency scores in automated spoken language tests.

Attacker: A test-taker attempting to manipulate the scoring system.

Application: The user prepends the adversarial segment to their spoken answer before submission. The model's transcription is distorted in a way that leads downstream scoring systems to output inflated proficiency levels (e.g., CEFR "C2").

Deployment: This attack can be performed by editing digital files before upload, or via over-the-air injection—e.g., playing the adversarial segment on a speaker before speaking into a microphone.

Note: In the over-the-air setting, environmental noise may distort the adversarial segment. To assess its robustness, we conducted a preliminary test where Gaussian white noise was added to the 3.2s segment used for attacking Qwen2-Audio (Table

Setup	∅	asl	WER
Without noise	99.9	0.0	100.0
With noise	99.6	0.0	100.0

Table 14: Experimental results of muting Qwen2-Audio when Gaussian white noise (amplitude range [-0.05, 0.05]) is added to the 3.2s learned adversarial segment.

2). As shown in Table 14, the perturbed segment retained its ability to mute the model, indicating that the adversarial effect is resilient to realistic distortions rather than being a fragile artifact.